Topical interests and the mitigation of search engine bias

S. Fortunato*†‡, A. Flammini*, F. Menczer*§¶, and A. Vespignani*‡

*School of Informatics, Indiana University, Bloomington, IN 47406; †Fakultät für Physik, Universität Bielefeld, D-33501 Bielefeld, Germany; §Department of Computer Science, Indiana University, Bloomington, IN 47405; and ‡Complex Networks Lagrange Laboratory, Institute for Scientific Interchange, 10133 Torino, Italy

Communicated by Elinor Ostrom, Indiana University, Bloomington, IN, July 1, 2006 (received for review March 2, 2006)

Search engines have become key media for our scientific, economic, and social activities by enabling people to access information on the web despite its size and complexity. On the down side, search engines bias the traffic of users according to their page ranking strategies, and it has been argued that they create a vicious cycle that amplifies the dominance of established and already popular sites. This bias could lead to a dangerous monopoly of information. We show that, contrary to intuition, empirical data do not support this conclusion; popular sites receive far less traffic than predicted. We discuss a model that accurately predicts traffic data patterns by taking into consideration the topical interests of users and their searching behavior in addition to the way search engines rank pages. The heterogeneity of user interests explains the observed mitigation of search engines' popularity bias.

PageRank | popularity bias | traffic | web graph

The topology of the Web as a complex, scale-free network is now well characterized (1–5). Several growth and navigation models have been proposed to explain the Web's emergent topological characteristics and their effect on users' surfing behavior (6–12). As the size and complexity of the Web have increased, users have become reliant on search engines (13, 14), so that the paradigm of search is replacing that of navigation as the main interface between people and the Web.‖ This trend leads to questions about the role of search engines in shaping the use and evolution of the Web.

A key assumption in understanding web growth is that pages attract new links proportionally to their popularity, measured in terms of traffic. According to preferential attachment and copy models (2, 6, 7), which explain the rich-get-richer dynamics observed in the Web's network structure, the traffic to each page is implicitly considered a linear function of the number of hyperlinks pointing to that page. The proportionality between popularity and degree is justified in a scenario in which two key assumptions hold. First, that pages are discovered and visited by users with a random web-surfing process; indeed this is the process modeled by the PageRank algorithm (see the supporting information, which is published on the PNAS web site). Second, PageRank, the likelihood of visiting a page, is linearly related to degree on average, which is supported by empirical data discussed in *Materials and Methods*. The use of search engines changes this scenario, mediating the discovery of pages by users with a combination of crawling, retrieval, and ranking algorithms that is believed to bias traffic toward popular sites. Pages highly ranked by search engines are more likely to be discovered by users and consequently linked from other pages. Because search engines heavily rely on link information to rank results, this would in turn increase the popularity of those pages even further. As popular pages become more and more popular, new pages would be unlikely to be discovered (14, 16). Such a vicious cycle (see the supporting information) would accelerate the feedback loop between popularity and number of links, introducing a nonlinear acquisition rate that would dramatically change the structure of the web graph from the current scale-free topology to a star-like network, where a set of sites would monopolize all traffic (17). The presumed popularity bias phenomenon (also known as "googlearchy") has been widely discussed in the computer, social, and political sciences (16, 18–22, **).

This paper offers an empirical study of the effect of search engines on the popularity of web pages by providing a quantitative analysis of the relationship between traffic and degree. We show that, contrary to common belief, the net popularity bias of search engines is much weaker than predicted in the literature. Even compared with the case in which no search occurs and all traffic is generated by surfing hyperlinks, search engines direct less traffic toward highly linked pages. More precisely, by empirical measurement in a large sample of web sites, we find a sublinear growth of traffic with in-degree. To explain this result, we refine a theoretical model of how users search and navigate the Web (16) by incorporating a crucial ingredient: the topical content of user queries. Such a realistic element reverses prior conclusions and accurately predicts the empirical relationship between traffic and in-degree. This finding suggests that a key factor in explaining the effect of search engines is the diversity and specificity of information sought by web users, as revealed by the wide variation of result samples matching user queries. In other words, search engines partially mitigate the rich-get-richer nature of the Web and give new sites an increased chance of being discovered, as long as they are about specific topics that match the interests of users. These results are important both in practical terms, for a quantitative assessment of page popularity, and conceptually, as a starting point for web growth models taking into account the interaction among search engines, user behavior, and information diversity.

Results and Discussion

For a quantitative definition of popularity we turn to the probability that a generic user clicks on a link leading to a specific page (21). We will also refer to this quantity as the traffic to the same page.

As a baseline to gauge the popularity bias of search, one can consider how web pages would gain popularity in the absence of search engines. People would browse web pages primarily by following hyperlinks. Other ways to discover pages, such as referral by friends, also would be conditional on a page being discovered by someone in the first place through links. To a first approximation, the amount of such surfing-generated traffic directed toward a given page is proportional to the number of links k pointing to it (in-degree). The more pages there are that point to that page, the larger the probability that a randomly surfing user will discover it. Successful second-generation search engines, Google being the premier example (23), have refined and exploited this effect in their

Conflict of interest statement: No conflicts declared.

¶To whom correspondence should be addressed. E-mail: fil@indiana.edu.

‖According to the Search Engine Round Table blog, WebSideStory Vice President Jay McCarthy announced at a 2005 Search Engine Strategies Conference that the number of page referrals from search engines had surpassed those from other pages. A more conservative estimate was obtained by monitoring web requests from a computer science department (15).

**Hindman, M., Tsioutsiouliklis, K. & Johnson, J. A., Annual Meeting of the Midwest Political Science Association, April 3–6, 2003, Chicago, IL.

 www.pnas.org/cgi/doi/10.1073/pnas.0605525103

ranking functions to gauge page importance. The PageRank value $p(i)$ of page i is defined as the probability that a random walker surfing the web graph will visit i next (see the supporting information), thereby estimating the page's discovery probability according to the global structure of the Web. Experimental observations and theoretical results show that, with good approximation, $p \sim k$ (see *Materials and Methods*). Therefore, in the absence of search engines, surfing traffic through a page would scale as $t \sim p \sim k$.

An alternative baseline to gauge the popularity bias of search would be the case of first-generation search engines. Because these engines did not use the topology of the web graph in their ranking algorithms, we would not expect the traffic generated by them to depend on in-degree. We focus on the former surfing baseline, which has been used as a reference in the literature (16) because it corresponds to the process modeled by Google's PageRank. The linear relationship between in-degree and traffic predicted by the surfing model is our benchmark against which the signature of search bias is characterized. As we show below, the intuitive googlearchy argument leads to a search model that predicts a superlinear relationship between t and k.

Modeling the Vicious Cycle. When navigation is mediated by search engines, to estimate the traffic directed toward a page, one must consider how search engines retrieve and rank results and how people use these results. According to Cho and Roy's approach (16), we need to find two relationships: (*i*) how the PageRank translates into the rank of a result page, and (*ii*) how the rank of a hit translates into the probability that the user clicks on the corresponding link, thus visiting the page.

The first step is to determine the scaling relationship between PageRank (and equivalently in-degree as discussed above) and rank. Search engines employ many factors to rank pages. Such factors are typically query-dependent: whether the query terms appear in the title or body of a page, for example. Search engines also use global (query-independent) importance measures, such as PageRank, to judge the value of search hits. Given that the true ranking algorithms used by commercial search engines are not public and that, for the sake of a simple model of global web traffic, we make the simplifying assumption that PageRank determines the average rank r of each page within search results, the page with the largest p has average rank $r \simeq 1$ and so on, in decreasing order of p.

To derive the relationship between p and r, we fitted the empirical curve of rank vs. PageRank obtained from a large WebBase crawl. We obtained a power law over three orders of magnitude:

$$r(p) \sim p^{-\beta}, \qquad [1]$$

where $\beta \approx 1.1$ (Fig. 1*A*). Cho and Roy (16) report a somewhat different value for the exponent β of 3/2. It is possible to obtain a larger exponent by fitting the tail of the curve, corresponding to high PageRank; however, we focused on the great majority of pages. To check this discrepancy, we used an alternative approach to study the relationship between r and p, illustrated in Fig. 1*B*, which confirmed our estimate of $\beta \approx 1.1$.

The second step is to approximate the traffic to a given page by the probability that, when the page is returned by a search engine, the user will click on its link. We expect the traffic t to a page to be a decreasing function of its rank r. Lempel and Moran (24) reported a nonlinear relationship confirmed by our analysis through query logs from AltaVista, as shown in Fig. 2. The data can be fitted quite well by a simple power-law relationship between the probability t that a user clicks on a hit and the rank r of the hit:

$$t \sim r^{-\alpha}, \qquad [2]$$

with exponent $\alpha \approx 1.6$. The fit exponent obtained by Cho and Roy (16) was 3/2, which is close to our estimate. The decrease

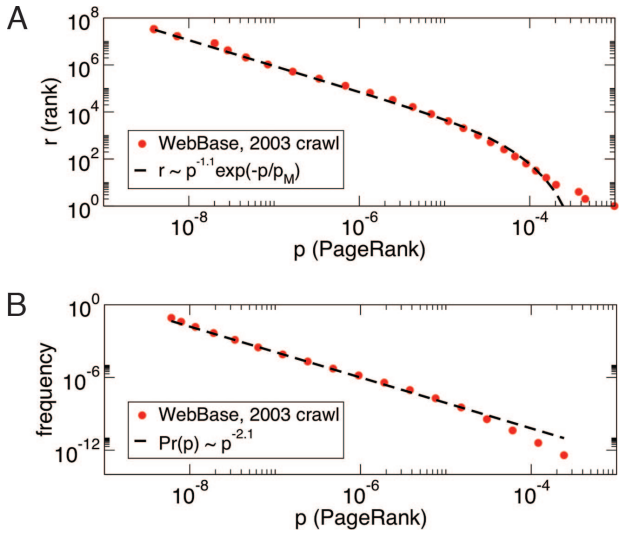

Fig. 1. Relationship between rank and PageRank. (*A*) Empirical relationship. The logarithm–logarithm plot shows a power law $r \sim p^{-1.1}$, with an exponential cutoff. (*B*) Distribution of PageRank p. The logarithm–logarithm plot shows a power law $\Pr(p) \sim p^{-2.1}$. The rank r is essentially the number of measures greater than p, i.e., $r = N \int_p^{p_{max}} Pr(x)dx$, where p_{max} is the largest measure gathered and N is the number of measures. In general, when the variable p is distributed according to a power law with exponent $-\mu$ and neglecting large N corrections, one obtains $r(p) \approx p^{1-\mu}$; therefore $\beta = \mu - 1 \approx 1.1$. Both plots are based on data from a WebBase 2003 crawl (http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase).

of t with the rank r of the hit clearly indicates that users focus with larger probability on the top results.

We are now ready to express the traffic as a function of page in-degree k by using a general scaling relationship, $t \sim k^{\gamma}$. The baseline (pure surfing model) is $\gamma = 1$; in the searching model, we take advantage of the relationships between t and r, between r and p, and between p and k to obtain

$$t \sim r^{-\alpha} \sim (p^{-\beta})^{-\alpha} = p^{\alpha\beta} \sim k^{\alpha\beta}. \qquad [3]$$

Therefore, $\gamma = \alpha \cdot \beta$, ranging between $\gamma \approx 1.8$ (according to our measures $\alpha \approx 1.6$, $\beta \approx 1.1$) and 2.25 (according to ref. 16).

In all cases, the searching model leads to a value $\gamma > 1$. This superlinear behavior is a quantitative prediction that corresponds to the presumed bias of search engines toward already popular sites. In this view, pages highly ranked by search engines are more likely to be discovered (as compared to pure surfing) and consequently linked-to by other pages, as shown empirically in ref. 16, which, in turn, would further increase their PageRank and raise the average

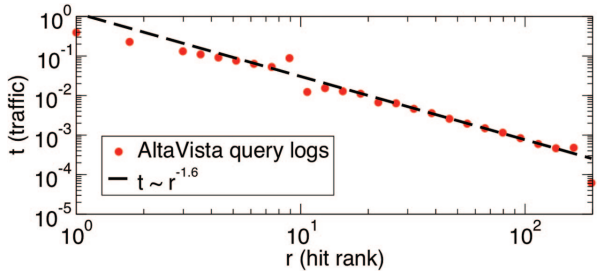

Fig. 2. Scaling relationship between click probability t and hit rank r. The logarithm–logarithm plot obtained with logarithmic binning of rank shows a power law with exponent $\alpha = 1.6 \pm 0.1$ (data from a sample of 7 million queries submitted to AltaVista between September 28, 2001 and October 3, 2001).

rank of those pages. Popular pages become more and more popular, whereas new pages are unlikely to be discovered. Such a vicious cycle would accelerate the rich-get-richer dynamics already observed in the Web's network structure (2, 6, 7). This presumed popularity bias or entrenchment effect has been recently brought to the attention of the technical web community (16, 20, 22), and methods to counteract it have been proposed (21, 22). There are also notable social and political implications to such a googlearchy (18, 19, **).

Empirical Data. To our knowledge, no prior empirical evidence exists to quantitatively support the vicious cycle theory by cross-correlating traffic with PageRank or in-degree data. Here we outline our effort to fill this void. Given a web page, its in-degree is the number of links pointing to it, which can be estimated by using search services offered by Google or Yahoo. Traffic is the fraction of all user clicks in some period that lead to each page; this quantity, also known as view popularity (21), is difficult to collect because search engines and Internet service providers protect their data for privacy and business reasons. To overcome this obstacle, we turned to the Alexa Traffic Rankings service, which monitors the sites viewed by users of its toolbar. Although our sources provide the best publicly available data for in-degree and traffic, there are some caveats on their use and reliability that are discussed in *Materials and Methods*. We used the Yahoo and Alexa services to estimate in-degree and traffic for a total of 28,164 web pages. Of these pages, 26,124 were randomly selected by using Yahoo's random page service. The remaining 2,040 pages were selected among the sites with the highest traffic. The resulting density plot is shown in Fig. 3*A*.

To derive a meaningful scaling relationship given the broad fluctuations in the data, we average traffic along logarithmic bins for in-degree, as shown in Fig. 3*B*. Surprisingly, both the searching and surfing models fail to match the observed scaling, which is not well modeled by a power law. Contrary to our expectation, the scaling relationship is sublinear; the traffic pattern is more egalitarian than what one would predict based on the simple search model described above or compared with the baseline model without search. Less traffic than expected is directed to highly linked sites. This finding suggests that some other factor must be at play in the behavior of web users, counteracting the skewed distribution of links in the Web and directing some traffic toward sites that users would never visit otherwise. Here, we revise the search model by taking into account the fact that users submit specific queries about their interests. This crucial element was neglected in the simple search model and offers a compelling interpretation of the empirical data.

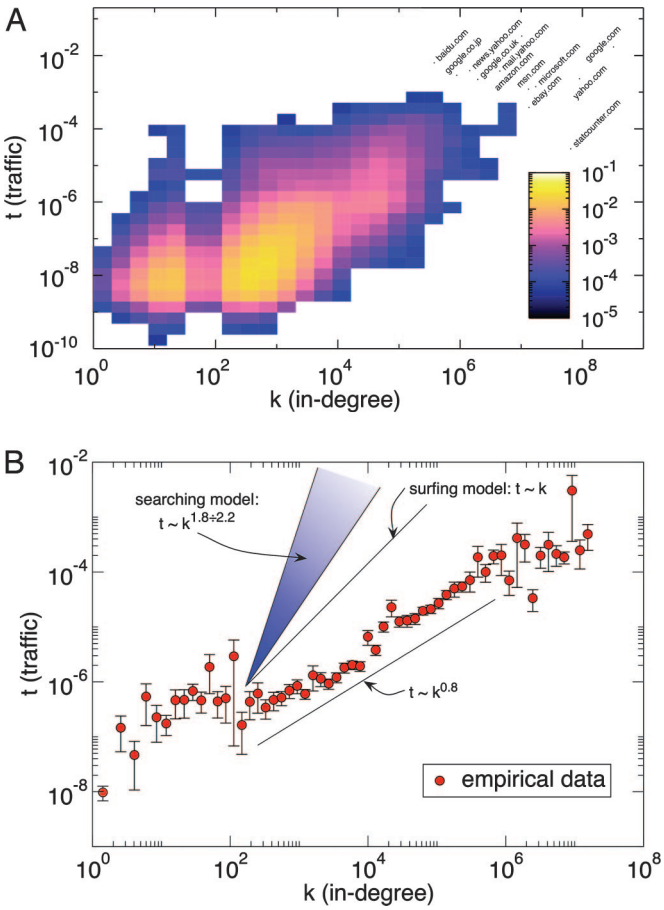

Fig. 3. Relationship between traffic and in-degree. (*A*) Density plot of traffic vs. in-degree for a sample of 28,164 web sites. Colors represent the fraction of sites in each log-size bin on a logarithmic color scale. A few sites with highest in-degree and/or traffic are highlighted. The source of in-degree data is Yahoo; using Google yields the same trend. Traffic is measured as the fraction of all page views in a 3-month period, according to Alexa data. The density plot highlights broad fluctuations in the data. (*B*) Relationship between average traffic and in-degree obtained with logarithmic binning of in-degree. Error bars correspond to ±1 SE. The power-law predictions of the surfing and searching models discussed in the text also are shown, together with a guide to the eye for the portion of the empirical traffic curve that can be fitted by a power law $t \sim k^{\gamma}$ ($\gamma \approx 0.8$).

Incorporating User Interests into the Search Model. In the previous theoretical estimate of traffic as driven by search engines, we considered the global rank of a page, computed across all pages indexed by the search engine. However, any given query typically returns only a small number of pages compared with the total number indexed by the search engine. The size of the "hit" set and the nature of the query introduce a significant bias in the sampling process. If only a small fraction of pages are returned in response to a query, their rank within the set is not representative of their global rank as induced, say, by PageRank.

To illustrate the effect of hit set size, let us assume that query result lists derive from a Bernoulli process such that the number of hits relevant to each query is on average $h \cdot N$, where h is the relative hit set size. In *Materials and Methods*, we show that this assumption leads to an alteration in the relationship between traffic and in-degree. Fig. 4*A* shows how the click probability changes with h. The result, $t \sim k^{\gamma}$, holds only in the limit case $h \to 1$. Because the size of the hit sets is not fixed but depends on user queries, we measured the distribution of hit set sizes for actual user queries as shown in Fig. 4*B*, yielding $\Pr(h) \sim h^{-\delta}$, with $\delta \approx 1.1$ over seven orders of magnitude. The exponential cutoff in the distribution of h is due to the maximum size of actual hit lists corresponding to non-noise terms and can be disregarded for our analysis (see the supporting information).

The traffic behavior is therefore a convolution of the different curves reported in Fig. 4*A*, weighted by $\Pr(h)$. The final relationship between traffic and degree can thus be obtained by numerical techniques. Strikingly, the resulting behavior reproduces the empirical data over four orders of magnitude, including the peculiar saturation observed for high-traffic sites (Fig. 4*C*). In *Materials and Methods*, we discuss the simulation and fitting techniques, as well as the trend in the low in-degree portion of the empirical curve.

The search model that accounts for user queries predicts a traffic trend for pages with increasing in-degree that is noticeably slower than the predictions of both the surfing model (baseline) and the naive searching model. The new element in the model is the simple fact that user interests tend to be specific, providing low-degree pages with increased visibility when they match user queries. In other words, the combination of search engines, semantic attributes of queries, and users' own behavior provides us with a compelling

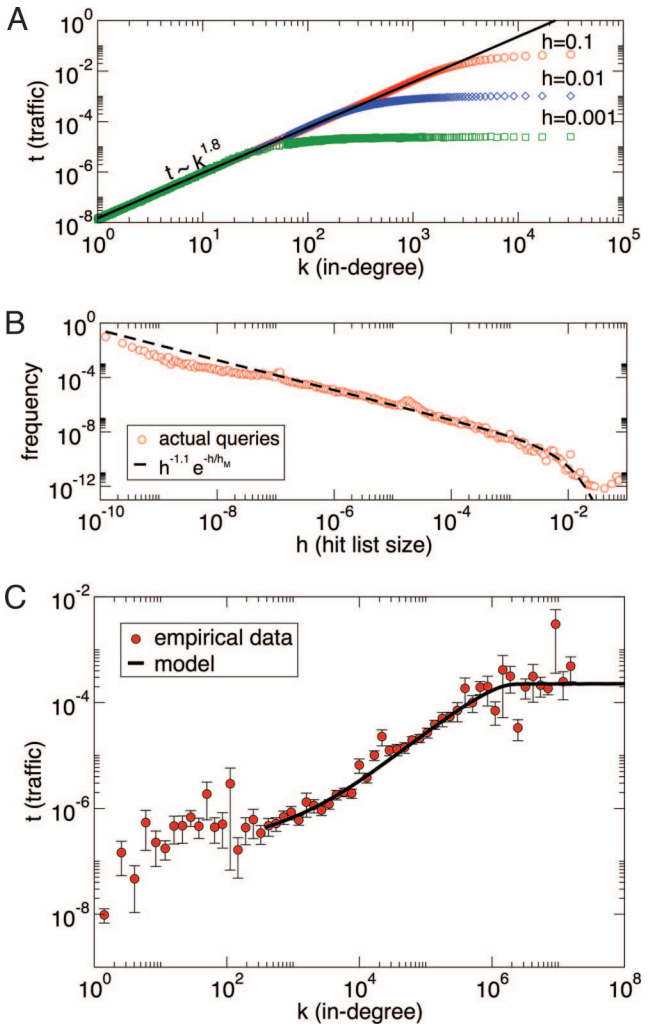

Fig. 4. Relationship between traffic, in-degree, and hit set size. (*A*) Scaling relationship between traffic and in-degree when each page has a fixed probability h of being returned in response to a query. The curves (not normalized for visualization purposes) are obtained by simulating the process $t[r(k), h]$ (see *Materials and Methods*). (*B*) Distribution of relative hit set size h for 200,000 actual user queries from AltaVista logs. The hit set size data were obtained from Google. Frequencies are normalized by logarithmic bin size. The logarithm–logarithm plot shows a power law with an exponential cutoff. (*C*) Scaling between traffic and in-degree obtained by simulating 4.5 million queries with a realistic distribution of hit set size. Empirical data are as shown in Fig. 3*B*. The trend in the low-k region can also be recovered (see *Materials and Methods*).

interpretation of how the rich-get-richer dynamics of the Web is mitigated by the search process.

Of course, actual web traffic is the result of both surfing and searching behaviors. Users rely on search engines heavily but also navigate from page to page through static links as they explore the neighborhoods of pages returned in response to search queries (15). It would be easy to model a mix of our revised searching model with the random surfing behavior. The resulting mixture model would yield a prediction somewhere between the linear scaling $t \sim k$ of the surfing model (compare with Fig. 3*B*) and the sublinear scaling of our searching model (compare with Fig. 4*C*). The final curve would still be sublinear, in agreement with the empirical traffic data. Users may also end up on a page by other mechanisms, such as bookmarks or email referrals. The simple models presented here neglect these mechanisms.

Conclusions

Our heavy reliance on search engines as a means of coping with the Web's size and growth does affect how we discover, visit, and link pages. Yet, despite the rich-get-richer dynamics implicit in the link analysis used to rank results, the use of search engines appears to mitigate the average traffic attraction of high-degree pages. The sublinear scaling relationship between traffic and page in-degree, revealed by our empirical measurements, is consistent with the observation that search engines lead users to visit ≈20% more pages than surfing alone (15). Such an effect may be understood within our theoretical model of search that considers the users' clicking behavior, the ranking algorithms used by search engines, and the long-tailed distribution observed for the number of hits matching user queries.

There are other possible interpretations for the sublinear scaling observed in the data. For instance, the quality of search engines might decrease the motivation for linking to already popular sites, whereas people may feel more motivated to link pages that do not appear among the top hits returned by search engines. Our search model, however, presents a very compelling explanation of the data because it predicts the traffic trend so accurately using a minimal account of query content and making strong simplifying assumptions, such as the use of PageRank as the sole ranking factor.

Our result has relevant conceptual and practical consequences; it suggests that, contrary to intuition and prior hypotheses, the use of search engines contributes to a more level playing field in which new sites have a greater chance of being discovered and thus of acquiring links and popularity, as long as they are about specific topics that match the interests of users as expressed through their search queries.

Such a finding is particularly relevant for the design of realistic models for web growth. The connection between the popularity of a page and its acquisition of new links has led to the well known rich-get-richer growth paradigm that explains many of the observed topological features of the Web. The present findings, however, show that several nonlinear mechanisms involving search engine algorithms and user behavior regulate the popularity of pages. A theoretical framework must consider more of the various behavioral and semantic issues that shape the evolution of the Web. Our current theoretical effort is to study how such a framework may yield coherent models that still agree with the Web's observed topological properties (25).

Finally, the present results provide a quantitative estimate of, and prediction for, the popularity and traffic generated by web pages. This estimate promises to become an important tool to be exploited in the optimization of marketing campaigns, the generation of traffic forecasts, and the design of future search engines.

Materials and Methods

Relationship Between In-Degree and PageRank. A mean field analysis has shown that in a directed network there is a precise relationship between a given in-degree k and the average PageRank p of all of the nodes with in-degree k (26). In the case of the web graph, owing to weak degree–degree correlations, this relationship is well approximated by a simple proportionality. To illustrate such behavior, we carried out a numerical analysis of PageRank on two web crawls performed in 2001 and 2003 by the WebBase collaboration at Stanford. The graphs are quite large: The former crawl has 80,571,247 pages and 752,527,660 links, and the latter crawl has 49,296,313 pages and 1,185,396,953 links. In our calculations of PageRank, we used a damping factor of 0.85, as in the original version of the algorithm (23) and many successive studies.

In Fig. 5, we averaged the PageRank values over logarithmic bins of in-degree. The data points mostly fall on a power-law curve for both samples, with p increasing with k. The estimated exponents of the power-law fits for the two curves are 1.1 ± 0.1 (2001) and 0.9 ± 0.1 (2003). The two estimates are compatible with the linear

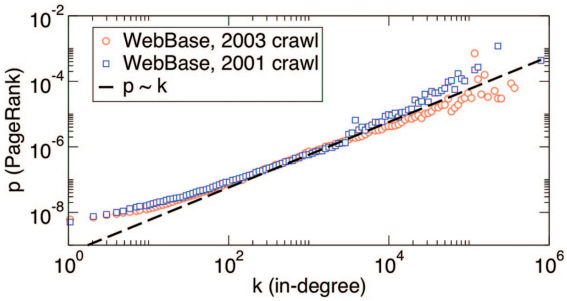

Fig. 5. PageRank as a function of in-degree for two samples of the Web taken in 2001 and 2003.

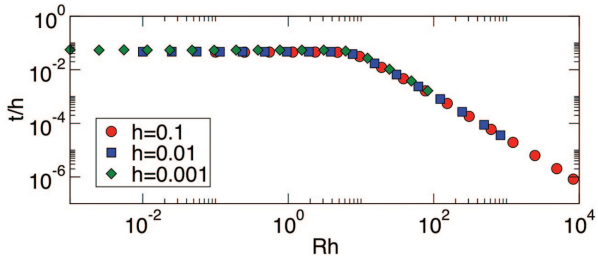

Fig. 6. Scaling of $t(R, N, h)/h$ with the variable $R{\cdot}h$. The three curves refer to a sample of $N = 10^5$ pages.

prediction between PageRank and in-degree used in our model. However, PageRank values may fluctuate considerably among pages with equal in-degree, consistent with the broad range of correlation coefficients between k and p reported in the literature (26–28).

For large values of in-degree, slight deviations from the linear prediction are observed in the crawl data of Fig. 5. Such deviations do not play a role in the relationship between in-degree and traffic because of the saturation of traffic in that region observed in the empirical data (Fig. 4*C*). If traffic is approximately independent of in-degree in this region, it is also independent of PageRank irrespective of the exact relationship between k and p.

Measuring In-Degree and Traffic. To ensure the robustness of our analysis, we collected our in-degree data twice at a distance of 2 months. Although there were differences in the numbers (for example, Yahoo increased the size of its index significantly in the meantime), there were no differences in the scaling relationships. We also collected in-degree data with Google, again yielding different numbers but the same trend. The in-degree measures exclude links from the same site. For example, to find the in-degree for http://informatics.indiana.edu, we would submit "link:http://informatics.indiana.edu/-site:informatics.indiana.edu" as the query. Note that the in-degree data provided by search engines are 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 in-links use various unpublished approximations based on sampling.

Alexa collects and aggregates historical traffic data from millions of Alexa Toolbar users. Although this is the only public source of web traffic data, it is generated by a sample of the web population that may be biased. Traffic is measured as page views per million in a 3-month period. Multiple page views of the same page made by the same user on the same day are counted only once. Our measure of traffic t corresponds to Alexa's count, divided by 10^6 to express the fraction of all of the page views by toolbar users that go to a particular site. Because traffic data are only available for web sites rather than single pages, we correlate the traffic of a site with the in-degree of its main page. For example, suppose that we want the traffic for http://informatics.indiana.edu. Alexa reports the 3-month average traffic of the domain indiana.edu as 9.1 page views per million. Furthermore, Alexa reports that 2% of the page views in this domain go to the informatics.indiana.edu subdomain. Thus, we reach the estimate of 0.182 page views per million, $t = 1.82 \times 10^{-7}$. The estimate of traffic by domains rather than pages introduces a systematic bias by which page traffic is overestimated. However, there is no reason to suspect that such a bias is correlated with degree. Therefore, it should be of no consequence for the exponent describing the relationship between degree and traffic.

Simulation of Search-Driven Web Traffic. When a user submits a query to a search engine, the latter will select all pages deemed relevant from its index and display the corresponding links ranked according to a combination of query-dependent factors, such as the similarity between the terms in the query and those in the page, and query-independent prestige factors, such as PageRank. Here we focus on PageRank as the main global ranking factor, assuming that query-dependent factors are averaged out across queries. The number of hit results depends on the query, and it is in general much smaller than the total number of pages indexed by the search engine.

Let us start from the relationship between click probability and rank in Eq. **2**. If all N pages in the index were listed in each query, as implicitly assumed in ref. 16, the probability for the page with the smallest PageRank to be clicked would be N^{α} ($\alpha \approx 1.6$ in our study) times smaller than the probability to click on the page with the largest PageRank. If instead, both pages ranked first and Nth appear among the n hits of a realistic query (with $n \ll N$), they would still occupy the first and last positions of the hit list, but the ratio of their click probabilities would be much smaller than before, i.e., n^{α}. This effect leads to a redistribution of the clicking probability in favor of lower-ranked pages, which are then visited much more often than one would expect at first glance. To quantify this effect, we must first distinguish between the global rank induced by PageRank across all web pages and the query-dependent rank among the hits returned by the search engine in response to a particular query. Let us rank all N pages in decreasing order of PageRank, such that the global rank is $R = 1$ for the page with the largest PageRank, followed by $R = 2$ and so on.

Let us assume for the moment that all query result lists derive from a Bernoulli process with success probability h (i.e., the number of hits relevant to each query is on average $h{\cdot}N$). The assumption that each page can appear in the hit list with the same probability h is in general not true, because there are pages that are more likely to be relevant than others, depending on their size, intrinsic appeal, and so on. If one introduces a fitness parameter to modulate the probability for a page to be relevant with respect to a generic query, the results would be identical as long as the fitness is not correlated with the PageRank of the page. In what follows, we stick to the simple assumption of equiprobability.

The probability $\Pr(R, r, N, n, h)$ that the page with global rank R has rank r within a list of n hits is

$$\Pr(R, r, N, n, h) = h^n(1-h)^{N-n}\binom{R-1}{r-1}\binom{N-R}{n-r}. \quad [4]$$

The probability for the Rth page to be clicked is then

$$t(R, N, h) = \sum_{n=1}^{N}\sum_{r=1}^{n}\frac{r^{-\alpha}h^n(1-h)^{N-n}}{\sum_{m=1}^{n}m^{-\alpha}}\binom{R-1}{r-1}\binom{N-R}{n-r}, \quad [5]$$

where we summed over the possible ranks r of R in the hit list ($r \in 1 \ldots n$) and over all possible hit set sizes ($n \in 1 \ldots N$). The sum in the denominator ensures the proper normalization of the click probability within the hit list.

From Eq. **5**, we can see that if $h = 1$, which corresponds to a list with all N pages, one recovers Eq. **2**, as expected. For $h < 1$, however, it is not possible to derive a close expression for $t(R, N, h)$, so we performed Monte Carlo simulations of the process leading to Eq. **5**.

In each simulation, we produce a large number of hit lists, where every list is formed by picking each page of the sample with probability h. At the beginning of the simulation, we initialize all entries of the array $t(R, N, h) = 0$. Once a hit list is completed, we add to the entries of $t(R, N, h)$, corresponding to the pages of the hit list, the click probability as given by Eq. **2** (with the proper normalization). With this Monte Carlo method, we simulated systems with up to $N = 10^6$ items. To eliminate fluctuations, we averaged the click probability in logarithmic bins, as already done for the experimental data.

We found that the function $t(R, N, h)$ obeys a simple scaling law:

$$t(R, N, h) = hF(Rh)A(N), \qquad [6]$$

where $F(R{\cdot}h)$ has the following form:

$$F(Rh) \sim \begin{cases} const & \text{if } h \le Rh \le 1 \\ (Rh)^{-\alpha} & \text{if } Rh > 1. \end{cases} \qquad [7]$$

An immediate implication of Eq. **6** is that, if one plots $t(R, N, h)/h$ as a function of $R{\cdot}h$ for N fixed, one obtains the same curve $F(R{\cdot}h)A(N)$ independently of the value of h (Fig. 6).

The decreasing part of the curve, $t(R, N, h)$, for $R{\cdot}h > 1$, i.e., $R > 1/h$, is the same as in the case when $h = 1$ (Eq. **2**), which means that the finite size of the hit list affects only the top-ranked $1/h$ pages. The effect is thus strongest when the fraction h is small, i.e., for specific queries that return few hits. The striking feature of Eq. **7** is the plateau for all pages between the first and the $1/h$th, implying that the difference in the values of PageRank among the top $1/h$ pages does not produce a difference in the probability of clicking on those pages. For $h = 1/N$, which would correspond to lists containing on average a single hit, each of the N pages would have the same probability of being clicked, regardless of their PageRank.

So far, we assumed that the number of query results is drawn from a binomial distribution with a mean of $h{\cdot}N$ hits. On the other hand, we know that real queries generate a broad range of possible hit set sizes, going from lists with only a single result to lists containing tens of millions of results. If the size of the hit list is distributed according to some function $\Pr(h)$, one would need to convolve $t(R, N, h)$ with $\Pr(h)$ to get the corresponding click probability:

$$t(R, N) = \int_{h_m}^{h_M} \Pr(h)t(R, N, h)dh, \qquad [8]$$

where h_{m} and h_{M} are the minimal and maximal fraction of pages in a list, respectively. We stress that if there is a maximal hit list size $h_{\mathrm{M}} < 1$, the click probability $t(R, N, h)$ will be the same for the first $1/h_{\mathrm{M}}$ pages, independent of the distribution function $\Pr(h)$.

The functional form of the real hit list size distribution $\Pr(h)$ (compare with Fig. 4*B*) is discussed in the supporting information. As to the full shape of the curve $t(R, N)$ for the Web, we performed a simulation for a set of $N = 10^6$ pages. We used $h_{\mathrm{m}} = 1/N$ because there are hit lists with a few or even a single result. The size of our sample allowed us to predict the trend between traffic and in-degree over almost six orders of magnitude for k. To fit the empirical data, we note that the theoretical curves obey a simple scaling relationship. It is indeed possible to prove that $t(R, N)$ is a function of the "normalized" rank R/N (and of N) and not of the absolute rank R. As a consequence, by properly shifting curves obtained for different N values along logarithmic x and y axes, it is possible to make the curves overlap (see the supporting information), allowing us to safely extrapolate to much larger N and to lay the curve derived by our simulation on the empirical data (as we did in Fig. 4*C*). However, because of the difficulty to simulate systems with as many pages as the real Web ($N \approx 10^{10}$), we could not extend the prediction to the low in-degree portion of the empirical curve. Wanting to extend the simulation beyond a million nodes, one would have to take into account that the proportionality assumption between p and k is not valid for $k < 100$, as shown in Fig. 5. By considering the flattening of PageRank in this region, one could recover in our simulation the traffic trend for small degree in Fig. 4*C*.

We thank Junghoo Cho, the anonymous reviewers, and the members of the Networks and Agents Network at Indiana University for helpful feedback on early versions of the manuscript; Alexa, Yahoo, and Google for extensive use of their web services; the Stanford WebBase project for crawl data; and AltaVista for use of its query logs. This work was funded in part by a Volkswagen Foundation grant (to S.F.), by National Science Foundation Awards 0348940 (to F.M.) and 0513650 (to A.V.), and by the Indiana University School of Informatics.

1. Albert, R., Jeong, H. & Barabási, A.-L. (1999) *Nature* **401,** 130–131.
2. Kleinberg, J., Kumar, S., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999) *Lect. Notes Comput. Sci.* **1627,** 1–18.
3. Broder, A., Kumar, S., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000) *Comput. Networks* **33,** 309–320.
4. Adamic, L. & Huberman, B. (2000) *Science* **287,** 2115.
5. Kleinberg, J. & Lawrence, S. (2001) *Science* **294,** 1849–1850.
6. Barabási, A.-L. & Albert, R. (1999) *Science* **286,** 509–512.
7. Kumar, S., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. & Upfal, E. (2000) in *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science* (IEEE Comput. Soc., Silver Spring, MD), pp. 57–65.
8. Kleinberg, J. (2000) *Nature* **406,** 845.
9. Pennock, D., Flake, G., Lawrence, S., Glover, E. & Giles, C. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 5207–5211.
10. Menczer, F. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14014–14019.
11. Menczer, F. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 5261–5265.
12. Barrat, A., Barthelemy, M. & Vespignani, A. (2004) *Lect. Notes Comput. Sci.* **3243,** 56–67.
13. Lawrence, S. & Giles, C. (1998) *Science* **280,** 98–100.
14. Lawrence, S. & Giles, C. (1999) *Nature* **400,** 107–109.
15. Qiu, F., Liu, Z. & Cho, J. (2005) in *Proceedings of the Eighth International Workshop on the Web and Databases*, eds. Doan, A., Neven, F., McCann, R. & Bex, G. J. (Assoc. Comput. Mach., New York), pp. 103–108, available at http://webdb2005.uhasselt.be.
16. Cho, J. & Roy, S. (2004) in *Proceedings of the 13th International Conference on the World Wide Web*, eds. Feldman, S. I., Uretsky, M., Najork, M. & Wills, C. E. (Assoc. Comput. Mach., New York), pp. 20–29.
17. Krapivsky, P. L., Redner, S. & Leyvraz, F. (2000) *Phys. Rev. Lett.* **85,** 4629–4632.
18. Introna, L. & Nissenbaum, H. (2000) *IEEE Comput.* **33,** 54–62.
19. Mowshowitz, A. & Kawaguchi, A. (2002) *Commun. ACM* **45,** 56–60.
20. Baeza-Yates, R., Saint-Jean, F. & Castillo, C. (2002) *Lect. Notes Comput. Sci.* **2476,** 117–130.
21. Cho, J., Roy, S. & Adams, R. (2005) *Proceedings of the ACM International Conference on Management of Data* (Assoc. Comput. Mach., New York), pp. 551–562.
22. Pandey, S., Roy, S., Olston, C., Cho, J. & Chakrabarti, S. (2005) *Proceedings of the 31st International Conference on Very Large Databases*, eds. Böhm, K., Jensen, C. S., Haas, L. M., Kersten, M. L., Larson, P.-Å. & Ooi, B. C. (Assoc. Comput. Mach., New York), pp. 781–792.
23. Brin, S. & Page, L. (1998) *Comput. Networks* **30,** 107–117.
24. Lempel, R. & Moran, S. (2003) *Proceedings of the 12th International Conference on World Wide Web* (Assoc. Comput. Mach., New York), pp. 19–28.
25. Fortunato, S., Flammini, A. & Menczer, F. (2006) *Phys. Rev. Lett.* **96,** 218701.
26. Fortunato, S., Boguna, M., Flammini, A. & Menczer, F. (2005) arXiv: cs.IR/0511016.
27. Pandurangan, G., Raghavan, P. & Upfal, E. (2002) *Lect. Notes Comput. Sci.* **2387,** 330–339.
28. Donato, D., Laura, L., Leonardi, S. & Millozzi, S. (2004) *Eur. Phys. J. B* **38,** 239–243.