Commentary

Editor's Note: The misuse of journal impact factor in hiring and promotion decisions is a growing concern. This article is one in a series of invited commentaries in which authors discuss this problem and consider alternative measures of an individual's impact.

Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks

Sergei Maslov[1] **and Sidney Redner**[2]

[1]Department of Condensed Matter Physics and Materials Science, Brookhaven National Laboratory, Upton, New York 11973, and [2]Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215

The number of citations is the most commonly used metric for quantifying the importance of scientific publications. However, we all have anecdotal experiences that citations alone do not characterize the importance of a publication. Some of the shortcomings of using citations as a universal measure of importance include the following.

(1) It ignores the importance of citing papers: a citation from an obscure paper is given the same weight as a citation from a ground-breaking and highly cited work.

(2) The number of citations is ill suited to compare the impact of papers from different scientific fields. Due to factors such as size of a field and disparate citation practices, the average number of citations per paper varies widely between disciplines. An average paper is cited ~6 times in life sciences, 3 times in physics, and <1 times in mathematics.

(3) Many groundbreaking older articles are modestly cited due to a smaller scientific community when they were published. Furthermore, publications on significant discoveries often stop accruing citations once their results are incorporated into textbooks. Thus, citations consistently underestimate the importance of influential old papers.

These and related shortcomings of citation numbers are partially obviated by Google's PageRank algorithm (Brin and Page, 1998). As we shall discuss, PageRank gives higher weight to publications that are cited by important papers and also weights citations more highly from papers with few references. Because of these attributes, PageRank readily identifies a large number of scientific "gems": modestly cited articles that contain ground-breaking results.

In a recent study (Chen et al., 2007), we applied Google's PageRank to the citation network of the premier American Physical Society (APS) family of physics journals (*Physical Review A–E*, *Physical Review Letters*, *Reviews of Modern Physics*, and *Physical Review Special Topics*). Our study was based on all 353,268 articles published in APS journals since their inception in 1893 until June 2003 that have at least one citation from within this dataset. This set of articles has been cited a total of 3,110,839 times by other APS publications. Our study is restricted to internal citations—citations to APS articles from other APS articles. Other studies (Bollen et al., 2006; Bergstrom, 2007) use the PageRank algorithm on a coarse-grained scale of individual scientific journals to formulate an alternative to the Impact Factor.

We can think of the set of all APS articles and their citations as a network, with nodes representing articles and a directed link between two nodes representing a citation from a citing article to a cited article. In Google's PageRank algorithm (Brin and Page, 1998), a random surfer is initially placed at each node of this network and its position is updated as follows: (1) with probability $1 - d$, a surfer hops to a neighboring node by following a randomly selected outgoing link from the current node; (2) with probability d, a surfer "gets bored" and starts a new search from a randomly selected node in the entire network. This update is repeated until the number of surfers at each node reaches a steady value, the Google number. These nodal Google numbers are then sorted to determine the Google rank of each node.

The original Brin-Page PageRank algorithm used the parameter value $d = 0.15$ based on the observation that a typical web surfer follows of the order of six hyperlinks, corresponding to a boredom attrition factor $d = 1/6 \simeq 0.15$, before aborting and beginning a new search. The number of citation links followed by researchers is generally considerably <6. In Chen et al. (2007) and Walker et al. (2007), we argued that in the context of citations the appropriate choice is $d = 1/2$, corresponding to a citation chain to two links.

The Google number G and number of citations k to a paper are approximately proportional to each other for $k \gtrsim 50$ (Chen et al., 2007) so that, on average, they represent similar measures of importance (Fortunato et al., 2006a,b). The outliers with respect to this proportionality are of particular interest (Fig. 1). Although the top APS papers by PageRank are typically very well cited, some are modestly cited and quite old. The dispar-

Received Sept. 24, 2008; accepted Sept. 25, 2008.

Work by S.M. at Brookhaven National Laboratory was performed under contract DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy. S.R. gratefully acknowledges financial support from U.S. National Science Foundation Grant DMR0535503.

Correspondence should be addressed to Sidney Redner at the above address. E-mail: redner@bu.edu.

DOI:10.1523/JNEUROSCI.0002-08.2008

0270-6474/08/2811103-03$15.00/0

Table 1. Top non-review-article APS publications ranked according to PageRank (PR), citation number (CNR), and CiteRank (CR) (Walker et al., 2007)

PR	CNR	CR	Publication	Title	Author(s)
1	54	42	PRL 10, 531 (1963)	Unitary Symmetry and Leptonic Decays	Cabibbo (C)
2	5	10	PR 108, 1175 (1957)	Theory of Superconductivity	Bardeen, Cooper, and Schrieffer (BCS)
3	1	1	PR 140, A1133 (1965)	Self-Consistent Equations. . .	Kohn and Sham (KS)
4	2	2	PR 136, B864 (1964)	Inhomogeneous Electron Gas	Hohenberg and Kohn (HK)
5	6	58	PRL 19, 1264 (1967)	A Model of Leptons	Weinberg (W)
6	55	37	PR 65, 117 (1944)	Crystal Statistics. . .	Onsager (O)
7	95	293	PR 109, 193 (1958)	Theory of the Fermi Interaction	Feynman and Gell-Mann (FG)
8	17	13	PR 109, 1492 (1958)	Absence of Diffusion in. . .	Anderson (A)
9	1853	133	PR 34, 1293 (1929)	The Theory of Complex Spectra	Slater (S)
10	12	11	PRL 42, 673 (1979)	Scaling Theory of Localization	Abrahams, Anderson, et al. (AA)
11	712	106	PR 43, 804 (1933)	. . . Constitution of Metallic Sodium	Wigner and Seitz (WS)

The full database of PageRank and CiteRank values of APS publications is accessible at http://www.cmth.bnl.gov/∼maslov/citerank.

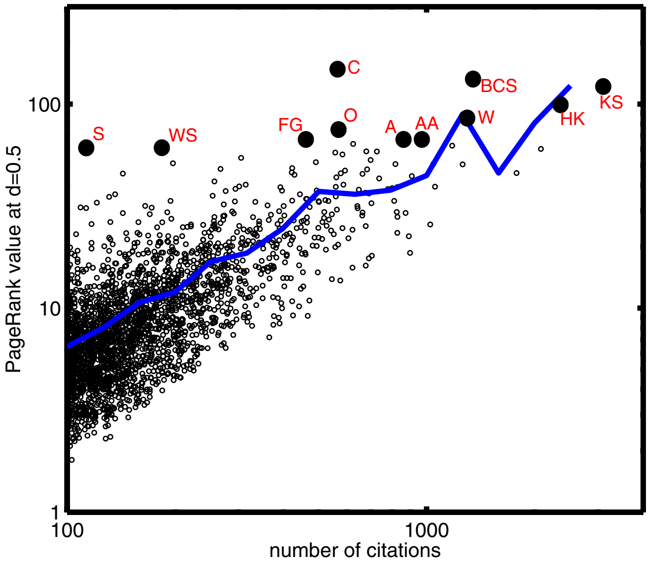

Figure 1. Scatter plot of the Google number versus number of citations for APS publications with >100 citations. The top Google-ranked papers (filled) are identified by author(s) initials and their details are given in Table 1. The solid curve is the average Google number of papers versus number of citations when binned logarithmically.

ity between PageRank and citation rank arises because the former involves both the number of citations and quality of the citing articles. The PageRank of a paper is enhanced when it is cited by its scientific "children" that themselves have high PageRank and when these children have short reference lists. That is, children should be influential and the initial paper should loom as an important "father figure" to its children.

The top articles according to Google's PageRank would be recognizable to almost all physicists, independent of their specialty (Table 1). For example, Onsager's (1944) exact solution of the two-dimensional Ising model (Fig. 1, the point labeled with *O*) was both a calculational tour de force as well as a central development in critical phenomena. The paper by Feynman and Gell-Mann (FG) introduced the $V - A$ theory of weak interactions that incorporated parity nonconservation and became the "standard model" of the field. Anderson's paper (A), "Absence of Diffusion in Certain Random Lattices" gave birth to the field of localization and was cited for the 1977 Nobel prize in physics. Particularly intriguing are the articles "The Theory of Complex Spectra," by J. C. Slater (S) and "On the Constitution of Metallic Sodium" by E. Wigner and F. Seitz (WS) with relatively few APS citations (114 and 184, respectively, as of June 2003). Slater's paper introduced the determinant form of the many-body wave-function that is so ubiquitous that the original work is no longer cited. Similarly, the latter paper introduced the Wigner–Seitz construction that constitutes an essential curriculum component in any solid-state physics course.

The PageRank algorithm was originally developed to rank web pages that are connected by hyperlinks and not papers connected by citations. The most important difference between these two networks is that, unlike hyperlinks, citations cannot be updated after publication. The constraint that a paper may only cite earlier works introduces a time ordering to the citation network topology. This ordering makes aging effects much more important in citation networks than in the World-Wide Web. The habits of scientists looking for relevant scientific literature are also different from web surfers. Apart from the already-mentioned shorter depth search $(1/d = 2)$, scientists often start literature searches with a paper of interest that they saw in a recent issue of a journal or heard presented at a conference. From this starting point a researcher typically follows chains citations that lead to progressively older publications.

These observations led us to modify the PageRank algorithm by initially distributing random surfers exponentially with age, in favor of more recent publications (Walker et al., 2007). This algorithm, CiteRank, is characterized by just two parameters: d (the inverse of the average citation depth) and τ (the time constant of the bias toward more recent publications at the start of searches). The optimal values of these parameters that best predict the rate at which publications acquire recent citations are $d = 0.5$ and $\tau = 2.6$ years for APS publications. With these values, the output of the CiteRank algorithm quantifies the relevance of a given publication in the context of currently popular research directions, whereas that of PageRank corresponds to its "lifetime achievement award." The database with the results of application of both algorithms to the citation network of APS publications can be accessed online at http://www.cmth.bnl.gov/∼maslov/citerank.

Google's PageRank algorithm and its modifications hold great promise for quantifying the impact of scientific publications. They provide a meaningful extension to traditionally used importance measures, such as the number of citations to articles and the impact factor for journals. PageRank implements, in a simple way, the logical notion that citations from more important publications should contribute more strongly to the rank of a cited paper. PageRank also effectively normalizes the impact of papers in different scientific communities (Xie et al., 2007). Other ways of attributing a quality for citations would require detailed contextual information about citation themselves, features that are presently unavailable. PageRank implicitly includes context by incorporating the importance of citing publications. Thus, PageRank represents a computationally simple and effective way to evaluate the relative importance of publications beyond simply counting citations.

We conclude with some caveats. It is very tempting to use citations and their refinements, such as PageRank, to quantify the importance of publications and scientists (Hirsch, 2005), especially as citation data becomes increasingly convenient to obtain electronically. In fact, the *h*-index of any scientist, a purported single-number measure of the impact of an individual scientific career, is now easily available from the Web of Science (http://apps.isiknowledge.com/). However, we must be vigilant for the overuse and misuse of such indices. All the citation measures devised thus far pertain to popularity rather than to the not-necessarily-coincident attribute of intrinsic intellectual value. Even if a way is devised to attach a high-fidelity quality measure to a citation, there is no substitute for scientific judgment to assess publications. We need to avoid falling into the trap of relying on automatically generated citation statistics for accessing the performance of individual researchers, departments, and scientific disciplines, and especially of allowing the evaluation task to be entrusted to administrators and bureaucrats (Adler et al., 2008).

References

Adler R, Ewing J, Taylor P (2008) Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS).

Bergstrom C (2007) Eigenfactor: measuring the value and prestige of scholarly journals. C&RL 68:314–316.

Bollen J, Rodriguez MA, Van de Sompel H (2006) Journal status. Scientometrics 69:669–687.

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30:107–117.

Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google. J Informetrics 1:8–15.

Fortunato S, Boguna M, Flammini A, Menczer F (2006a) How to make the Top Ten: Approximating PageRank from in-degree. In: Proceedings of the Fourth Workshop on Algorithms and Models for the Web-Graph (WAW2006), Banff, Alberta, Canada, November 30–December 1.

Fortunato S, Flammini A, Menczer F (2006b) Scale-free network growth by ranking. Phys Rev Lett 96:218701.

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci U S A 102:16569–19572.

Walker D, Xie H, Yan KK, Maslov S (2007) Ranking scientific publications using a model of network traffic. J Stat Mech 6:P06010.

Xie H, Yan KK, Maslov S (2007) Optimal ranking in networks with community structure. Physica A 373:831–836.