Assessing Computer Scientists Using Citation Data*

Glenn Ellison
MIT and NBER

April 2012

1 Introduction

In the last few years there has been a surge of interest in using bibliometric data to assess academics and academic departments. The field of computer science poses a particular challenge: publications have shifted from journals to conference proceedings. The recent National Research Council (2010) rankings reinforce the view that bibliometric analyses may be difficult: the NRC rankings are strikingly uninformative in some cases and reach dubious conclusions in others. For example, the NRC's "survey-based" ranking says only that MIT's program should rank between 2nd and 12th and it puts the traditional top-ten program at University of Washington somewhere between 17th and 55th.[1]

This paper investigates whether easy-to-collect Hirsch-like citation indexes can provide more compelling assessments of computer scientists and computer science departments and discusses how such assessments should be done. The paper follows the methodology that Ellison (2010) used to assess academic economists: the Hirsch-index is treated as just one of a large class of possible indexes and which index to use (and how to adjust for age differences and field-specific citation patterns) is assessed by considering which Hirsch-like indexes best predict which computer scientists are employed in which departments and which Hirsch-like indexes are most aligned with traditional rankings of computer science departments.

I find that variants of the Hirsch-index that place more weight on a smaller number of more highly cited papers are superior to the original Hirsch index. Appropriate citation indexes appear to be sufficiently powerful to provide fairly compelling departmental rankings: a simple linear combination of average citation indexes and department size is highly correlated with reputation-based rankings for highly-ranked U.S. departments. Whether

*Much of the work on this project was conducted while the author was a visiting researcher at Microsoft Research. I thank Hongkai Zhang for outstanding research assistance throughout the project and Matthias Beestermoeller for work in the initial stages. Financial support was provided by the NSF (SES-05508897), and the Toulouse Network for Information Technology.

[1] The University of Washington ranks 9th both in the 1993 NRC rankings and in the most recent US News & World Report rankings.

citation indexes can provide compelling rankings of individual computer scientists is less clear: many of the computer scientists with the highest indexes are ACM fellows, but there is a great deal of within-department variation.

2 Methodology

The Hirsch index of an academic researcher is defined to be the largest number h such that the researcher has authored (individually or jointly) at least h papers that have h or more citations each. As in Ellison (2010) this paper regards Hirsch's index as one possible choice from a broader class: the (a, b) *generalized Hirsch index* $h_{(a,b)}$ is defined to be the largest h such that the author has written at least h papers that have ah^b or more citations each.[2]

The first empirical exercise in this paper is an assessment of which indexes from this class are most appropriate for use in computer science. Indexes are evaluated in two ways. First, I consider how powerful each index is as a predictor of the department in which a computer scientist is tenured. Second, I consider how aligned indexes are with traditional reputational rankings of departments.

Te prediction model is that computer scientist i will work in department j with probability

$$p_{ij} = e^{\beta_0 \log(N_j) - \beta_1 d_{ij}} / \left(\sum_k e^{\beta_0 \log(N_k) - \beta_1 d_{ik}} \right), \tag{1}$$

where the latter sum is taken over all departments in the sample, N_j is the number of computer scientists in the sample used for this estimation who work in department j, and d_{ij} is the quality "mismatch" between computer scientist i and department j. The quality mismatch is assumed to take the form:

$$d_{ij} = (q_j - h_{(a,b)i} Z_i \gamma)^2.$$

The q_j are estimated parameters reflecting the unknown the quality of department j, Z_i is a vector of attributes of computer scientist i which will usually be taken to include years-post-Ph.D. and a set of control variables to correct for the subfields in which the researcher works. For each value of (a, b) considered, maximum likelihood estimation is used to estimate the parameters β, γ, and q, and the maximized log likelihood is used to assess the performace of the $h_{(a,b)}$ index. Note that this metric uses no information on department reputations: an index is regarded as performing well if departments appear to be collections of researchers with similar indexes.

[2] One other departure from Hirsch's methodology that will be maintained throughout this paper is that counts will be constructed so that authors get only $1/n$ credit for a paper with n-authors. The way in which this is done follows Egghe (2008) and Shreiber (2008). Ellison (2010) considers a broader class of indexes by viewing Hirsch's simple counts and $1/n$ weighted counts as two examples of a more general construction of giving $1/n^c$ credit for multiauthored papers.

The reputation-based criterion for assessing indexes involves regressing a traditional reputation-based measure of department quality on the average citation index of its faculty and the department size:

$$Reputation_j = \alpha_0 + \alpha_1 \text{Mean}_{i \in j} \left\{ h_{(a,b),i} e^{-Z_i \hat{\gamma}} \right\} + \alpha_2 FacultyCount_j + \epsilon_j. \tag{2}$$

Here, citation indexes are adjusted for field and experience differences using estimated correction factors from model (3) below. $FacultyCount_j$ is the number of faculty in a department eligible for inclusion in the sample.[3] It is included to reflect that reputational rankings of departments may reflect both average faculty quality and faculty size. Indexes are compared in terms of the R^2 of this regression.

The dataset used for the analysis contains information on 600 computer scientists from 29 U.S. departments. The sample consists of most computer scientists at these departments in the fall of 2009 who obtained Ph.D.'s from 1970 to 2000 and held the title of Associate Professor or Professor.[4] Citation counts were collected from Google Scholar for the first 100 papers listed there.[5] Researchers were assigned fractionally to thirteen subfields by looking for keywords in research descriptions posted on departmental websites. The measure of departmental reputation used is the simple arithmetic average of the 1993 NRC rating and the 2010 U.S. News and World report rating.

The remainder of the paper presents some additional analyses using the $h_{(30,1)}$ citation index. These are intended both to illustrate its power (and shortcomings) and to provide correction factors that could be used by researchers (or others) evaluating researchers using this citation index. Specifically, the paper estimates the model

$$\log(h_{(30,1),i}) = \beta_0 + \beta_1 Reputation_j + Z_i \gamma + \epsilon_i, \tag{3}$$

via an ordinary least squares regression, where $Reputation_j$ is again the reputational measure of department quality and the vector Z_i of control variables includes a quadratic in years-post-Ph.D. and the field corrections. The estimated experience and field coefficients provide the correction factors. The power of citation indexes is assessed by examining the goodness-of-fit of the regression and by examining lists of computer scientsts both informally and in terms of ACM Fellow status.

[3] This includes researchers dropped for any of the reasons noted in footnote 3.

[4] A total of 194 researchers were left out the sample. This included researchers with interests in biology (88 cases) or human-computer interaction (33) because they might publish in outlets not included in Google Scholar's "Engineering, Computer Science, and Mathematics" category, researchers with missing or hard-to-classify research interests (47), researchers with missing data on Ph.D. year (4), researchers listed in multiple departments (4), and researchers for whom name issues made collecting citations more difficult (18).

[5] Google Scholar citation counts have the advantage of covering recent papers and not being restricted to papers published in journals. They tend to be substantially higher than ISI counts. One limitation is that the display format Google used in the period when the data was collected leads to undercounting of the number of authors on many papers with more than three authors.

3 Results

3.1 Comparisons of various Hirsch-like indexes

Variants of the Hirsch index that use larger values of a and/or b can be thought of as emphasizing the citations to a smaller number of papers. The mean value of the $h_{(1,1)}$ index across the computer scientists in our sample is 21.5. (Recall that we are only giving partial credit for coauthored papers so a count of 21.5 may reflect 50 papers.) The $h_{(25,1)}$, $h_{(5,2)}$, and $h_{(1,3)}$ indexes all have means of about 5. The means of the $h_{(100,1)}$ and $h_{(40,2)}$ indexes are around 2.3.

The left panel of Table 1 reports the maximized per-obervation log likelihood from estimating the model (1) using various Hirsch-like indexes. Variants of the Hirsch index that use larger values of a and/or b provide better fits than the $h_{(1,1)}$ index. The best fits are obtained from indexes with $a = 100$. Many more moderate indexes provide about two-thirds of the improvement.[6]

The right panel of Table 1 reports the R^2 of regression (2) estimated using various $h_{(a,b)}$. Many $h_{(a,b)}$ provide substantially better fits than does the $h_{(1,1)}$. The best fits in each column are obtained when using the $h_{(10,1)}$, $h_{(2,2)}$, and $h_{(2,3)}$ indexes. Indexes with somewhat larger a provide only slightly worse fits.

Taken together, the two analyses strongly suggest that it is advisable to use indexes with larger a and/or b rather than the $h_{(1,1)}$ index. They provide somewhat different views, however, of which indexes are best. Given that many indexes perform similarly, there appears to be little reason to use indexes with $b > 1$ rather than of indexes with $b = 1$ (which are easier to compute mentally). In the analyses below I will use the $h_{(30,1)}$ index which seems to perform fairly well on both criteria. But other indexes would also have been reasonable choices.

3.2 How aligned are citation indexes with individual employment outcomes?

In this subsection and the one that follows I provide additional results on the relation between the $h_{30,1}$ index and employment outcomes and reputations. Table 2 presents regression coefficients from estimating the regression model (3). The coefficient on the department reputation variable is highly statistically significant showing that computer scientists at higher ranked departments clearly tend to have higher citation indexes. But the goodness-of-fit statistics show that there is a great deal of variation at the individual level. The R^2 of the regression is just 0.16. The root mean squared error of the model, 0.44, is roughly as large as the difference between the mean values of the index for computer scientists at the 1st and 20th ranked schools. This large residual error is due almost entirely to within-department

[6] Ellison (2010) found that the $h_{(5,2)}$ performed best in a similar analysis of academic economists.

	Max log likelihood of model (1) using $h_{(a,b)}$			R^2 of model (2) using $h_{(a,b)}$		
	Value of b			Value of b		
Value of a	1	2	3	1	2	3
0.5	-3.183	-3.166	-3.161	0.51	0.70	0.72
1	-3.177	-3.166	-3.160	0.58	0.72	0.70
2	-3.170	-3.161	-3.154	0.65	0.72	0.72
5	-3.162	-3.157	-3.152	0.69	0.72	0.72
10	-3.162	-3.152	-3.150	0.72	0.70	0.67
20	-3.155	-3.151	-3.154	0.71	0.70	0.65
30	-3.152	-3.148	-3.148	0.71	0.67	0.67
50	-3.147	-3.149	-3.147	0.69	0.65	0.68
100	-3.142	-3.143	-3.145	0.65	0.69	0.68
10,000	-3.156	-3.162	-3.159	0.55	0.63	0.66

Table 1: Goodness of fit for models using $h_{(a,b)}$ to explain employment outcomes and department reputations

variation: the root mean square error drops only slightly to 0.43 if one includes a dummy variable for each department. We must then conclude either that there is substantial heterogeneity in faculty quality within computer science departments and/or that the $h_{(30,1)}$ citation index is a very noisy measure of quality at the individual level.

The age coefficients in the regression indicate that citation indexes increase at about 1.3% per year for the youngest researchers in our sample (who are 10 years post-Ph.D.) and are highest for researchers who are about 40 years post-Ph.D. Presumably this reflects a balance of two main forces: older researchers have had longer to accumulate citations; but flow citations are much higher today than they were decades ago especially given that Google is more likely to have indexed more recent papers.

The coefficient estimates on the field effects provide correction factors that could be used when comparing researchers in different fields. Several of the estimated field effects are substantial. For example, the 0.27 coefficient for security and cryptography indicates that researchers in that field on average have citation indexes that are 31% higher than researchers in the omitted "other" category ($e^{0.27} \approx 1.31$). On the opposite side, researchers in scientific computing, theory, and programming languages have substantially lower citation indexes. Correcting for differences across fields appears desirable. Unfortunately, the estimates of the field effects are not highly precise.

To help readers get a sense for how well/poorly the indexes perform as a tool for ranking individuals Table 3 lists the four researchers with the highest field- and age-adjusted $h_{(30,1)}$ index at each school. In many cases the measure seems to do a good job of picking out faculty

Variable	Coefficient Estimate	Standard Error
Reputation	0.27	0.04
Years PostPhD	0.015	0.012
Years PostPhd^2	-0.00018	0.00026
Theory	-0.18	0.09
Programming languages	-0.19	0.12
Architecture	-0.12	0.10
Scientific computing	-0.29	0.15
Operating systems	0.03	0.19
Computer networks	0.11	0.12
Computer graphics	0.03	0.10
Security and cryptography	0.27	0.12
AI and machine learning	0.01	0.09
Databases and data mining	0.13	0.12
Robotics	0.15	0.12
Vision and speech	0.16	0.13
Constant	0.07	0.20

Table 2: Regression of individual $h_{(30,1)}$ on school reputation, age, and field effects

who are among the most highly regarded in their departments – 55 of the 116 researchers on the list are ACM fellows. In other cases, however, readers will probably also feel that the list illustrates the limitation of a simple citation-based measure. (It should be kept in mind that the list intentionally identifies researchers who are highly cited relative to their *age* and field. Without the age corrections the list would include more senior distinguished researchers. For example, the MIT list would become Rodney Brooks, Hari Balakrishnan, Tim Berners-Lee, and Berthold Horn.)

Table 3 also gives some sense of the correlation between how many highly cited researchers are in each department and traditional reputational rankings. (The departments are ordered by the *Reputation* variable.) The top three departments stand out as the only ones with at least four researchers with adjusted indexes over 8. The other two "top five" departments have at least four researchers over 7. Most of the 6th-10th ranked departments have their fourth-highest index in the 6's. Most of the 11th-20th ranked departments have their fourth-highest index in the 5's. And most of the remainder have their fourth-highest index in the 4's. But there are some substantial departures from this pattern. For example, UCLA, University of Massachusetts, and University of Pennsylvania have more highly cited researchers than other schools with similar reputational rankings.

	Faculty with highest field- and age-adjusted $h_{(30,1)}$ indexes							
Department	Researcher	$\tilde{h}_{(30,1)}$	Researcher	$\tilde{h}_{(30,1)}$	Researcher	$\tilde{h}_{(30,1)}$	Researcher	$\tilde{h}_{(30,1)}$
Stanford	H. Garcia-Molina	11.9	Sebastian Thrun	9.4	Jennifer Widom	9.0	Patrick Hanrahan	8.3
MIT	Hari Balakrishnan	10.2	Robert Morris	9.3	Rodney Brooks	9.2	David Karger	8.9
UC-Berkeley	Vern Paxson	10.7	Scott Shenker	10.5	David Culler	9.3	Randy Katz	8.5
Carnegie-Mellon	Randal Bryant	8.9	Takeo Kanade	8.8	Hui Zhang	7.4	Tuomas Sandholm	7.3
Cornell	Jon Kleinberg	9.2	Bart Selman	7.8	Dexter Kozen	7.7	Charles Van Loan	7.0
Princeton	Robert Tarjan	13.6	Robert Schapire	9.8	Bernard Chazelle	7.5	Kai Li	6.8
Illinois	Jiawei Han	9.0	Ralph Johnson	7.1	Jose Meseguer	6.6	William Gropp	6.3
Texas	Vladimir Lifschitz	8.3	E. Allen Emerson	7.6	Raymond Mooney	5.5	Kathryn McKinley	5.2
Washington	Pedro Domingos	8.1	Oren Etzioni	8.0	Dieter Fox	7.7	Dan Suciu	6.3
Wisconsin	Mark Hill	6.8	Thomas Reps	6.4	Gurindar Sohi	6.0	Susan Horwitz	6.0
Cal Tech	Yaser Abu-Mostafa	5.7	Mathieu Desbrun	5.5	Alan Barr	5.1	Steven Low	5.1
Harvard	Leslie Valiant	11.1	M. Mitzenmacher	7.8	H.T. Kung	6.5	Greg Morrisett	5.3
UC-Los Angeles	Deborah Estrin	9.6	D. Terzopoulos	9.6	Mario Gerla	7.9	Mani Srivastava	7.3
Maryland	William Pugh	6.4	Larry Davis	6.3	VS Subrahmanian	5.4	Hanan Samet	5.4
Michigan	Steven Abney	5.8	Kang Shin	5.8	Trevor Mudge	5.7	Peter Chen	5.3
Brown	John Hughes	6.6	Eugene Charniak	6.4	Maurice Herlihy	6.0	Michael Black	5.8
UC-San Diego	Yoav Freund	7.4	Y.Papakonstantinou	7.0	Mihir Bellare	6.6	George Varghese	6.5
Yale	Dana Angluin	7.9	Paul Hudak	7.8	Drew McDermott	6.2	David Gelernter	6.1
Georgia Tech	Greg Turk	6.8	Richard Lipton	6.5	Thad Starner	6.3	Wenke Lee	5.5
Columbia	M. Yannakakis	9.1	H. Schulzrinne	7.0	Luis Gravano	5.4	Salvatore Stolfo	5.2
Massachusetts	A. McCallum	9.4	W. Bruce Croft	8.4	Don Towsley	7.0	Jim Kurose	6.8
Rice	Moshe Vardi	7.9	Vivek Sarkar	5.8	Keith Cooper	5.1	J. Mellor-Crummey	4.7
Pennsylvania	Rajeev Alur	10.0	Fernando Pereira	8.0	Benjamin Pierce	6.8	Michael Kearns	6.8
Southern Calif.	Ramesh Govindan	7.4	Nenad Medvidovic	5.2	Shang-Hua Teng	4.7	Gerard Medioni	4.6
NYU	Richard Cole	6.3	Chee Yap	5.6	Victor Shoup	5.5	Margaret Wright	5.3
Purdue	Mikhail Atallah	4.9	Greg Frederickson	4.6	Eugene Spafford	4.5	Elisa Bertino	4.4
North Carolina	Ming Lin	6.1	James Anderson	5.2	Michael Reiter	5.1	Wei Wang	4.6
Duke	Donald Rose	6.1	Carlo Tomasi	5.4	Alvin Lebeck	4.6	Ronald Parr	4.1
Chicago	Ian Foster	12.9	Laszlo Babai	5.7	Partha Niyogi	5.1	David MacQueen	4.6

Table 3: Researchers with the highest field- and age-adjusted $h_{(30,1)}$ indexes by department

3.3 How aligned are citation-based and reputational rankings at the department level?

Estimating model (3) to explain department reputations as a function of department size and the mean adjusted citation index for the faculty gives:

$$\begin{array}{rcl} \textit{Reputation}_j & = & \underset{(0.31)}{1.63} + \underset{(0.07)}{0.42} \cdot \mathrm{Mean}_{i \in j} \left\{ h_{(30,1),i} e^{-Z_i \hat{\gamma}} \right\} + \underset{(0.004)}{0.017} \cdot N_j + \epsilon_j. \end{array}$$

The model turns out to place substantial weight on both the citation indexes and faculy size. The significance of the latter variable indicates that evaluators give higher marks to larger departments, which could reflect that they value the breadth and depth of field coverage that larger faculties provide.

Table 4 provides the implied bibliometric ranking of the 29 departments in the dataset. The bibliometric rating has an 0.84 correlation with the reputational rating (which again is a simple average of the 1993 NRC score and the current USNWR score). The three schools that are in an approximate tie for first in the reputational ranking (Stanford, MIT, Berkeley) occupy the top three spots in the bibliometric ranking with scores well above those of the other departments. The next three schools in the reputation ranking (Carnegie-Mellon, Cornell, Princeton) are also all in the top seven in the bibliometric ranking. The five departments with the lowest reputational ratings are all in the bottom eight in the bibliometric ranking. Some differences are that Georgia Tech, Yale, Massachusetts, and Pennslyvania look better in the bibliometric rating (the first largely due to its size) and that Cal Tech and Texas fare better in the reputational rating.

4 Conclusions

The first conclusion of this paper is that in contrast to what might be inferred from the most recent NRC ratings it seems quite possible to construct bibliometric rankings of computer science departments. Whereas the NRC's complex twenty-attribute model produced some rankings that are far from common perceptions, this paper shows that a very simple model using just citation data that can be quickly downloaded from Google Scholar and a measure of department size can align closely with traditional reputation-based rankings of the top departments. This suggests that citation-based indexes like those used here may be generally useful in assessing departments about which experts may be less well informed.

A second conclusion is that labor market outcomes for computer scientists appear to be more aligned with variants of Hirsch's index that focus on a smaller number of highly cited papers than they are with Hirsch's original index. The $h_{(30,1)}$ index which is very easy to compute seems like one fine choice, but many others could also be used.

A third conclusion is that it is not clear whether citation indexes like those used here can be developed into a compelling tool for evaluating individual computer scientists. The approach and sample used here provided only imprecisely estimated correction factors for field-specific differences in citations. And there appears to be substantial heterogeneity in the citation indexes of the faculty who work at each school. One potential explanation for this is that the indexes might be working well but there is only limited hierarchical sorting of faculty into departments, i.e. lower ranked departments have a number of extraordinary faculty and top departments have a number of faculty who are not so outstanding. But an alternate explanation is that the citation indexes may only provide quite noisy signals of how experts would view each individual. The noise could perhaps be reduced by improving the classification of individuals into fields. But it may be that the noise can mostly average out at the level of a department, but is inherently substantial at the individual level.

Elements of this paper's methodology may also be useful in future work. A plethora of bibliometric measures that could potentially be used to evaluate individuals and departments. Looking at how well measures align both with labor market outcomes and with reputation-based measures may be a useful way to select among them. And when selecting among or combining measures it may be useful to focus on the subsample of individuals or departments for which reputational measures are considered reliable.

References

Adler, Robert, John Ewing, and Peter Taylor (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)," mimeo.

Egghe, L. (2008): "Mathematical Theory of the h- and g-index in Case of Fractional Counting of Authorship," *Journal of the American Society of Information Science and Technology* 59 (10), 1608-1616.

Ellison, Glenn, (2010): "How Does the Market Use Citation Data? The Hirsch Index in Economics," National Bureau of Economic Research Working Paper 16419.

Hirsch, J. E. (2005): "An Index to Quantify an Individual's Scientific Research Output," *Proceedings of the National Academy of Sciences* 102, 16569-16572.

Kosmulski, Marek (2006): "A New Hirsch-type Index Saves Time and Works Equally Well as the Original h-Index ," ISSI Newsletter 2, 4-6.

National Research Council (2010): *A Data-Based Assessment of Research-Doctorate Programs in the United States.* Jeremiah P. Ostriker, Charlotte V. Kuh and James A. Voytuk (eds.). Washington D.C.: National Academies Press.

Schreiber, Michael (2008): "A Modification of the h-index: the h_m-index Accounts for Multi-authored Manuscripts," *Journal of Informetrics* 2 (3), 211-216.

Department	Bibliometric Rating	Reputation Rating	Average adj. $h_{(30,1)}$ index	Number of faculty in sample
MIT	5.05	4.96	5.98	54
UC-Berkeley	4.85	4.94	6.22	37
Stanford	4.85	4.99	6.17	38
Carnegie-Mellon	4.41	4.88	4.87	44
Princeton	4.35	4.36	5.70	20
Washington	4.34	4.27	5.10	34
Cornell	4.20	4.62	4.98	29
Georgia Tech	4.15	3.70	3.59	60
UC-Los Angeles	4.06	3.87	4.85	24
Illinois	4.05	4.35	4.33	36
Yale	4.01	3.72	5.15	14
Brown	3.95	3.78	4.96	15
Harvard	3.95	3.92	5.03	13
Massachusetts	3.90	3.65	4.30	28
Pennsylvania	3.90	3.61	4.62	20
UC-San Diego	3.90	3.73	4.29	28
Texas	3.88	4.29	3.96	35
Michigan	3.84	3.80	4.11	29
Columbia	3.83	3.68	4.37	22
Wisconsin	3.82	4.10	4.27	24
Maryland	3.81	3.85	3.80	35
Chicago	3.76	3.31	4.53	14
NYU	3.68	3.50	3.90	25
North Carolina	3.58	3.43	3.57	27
Rice	3.56	3.63	4.05	14
Southern Calif.	3.52	3.61	3.68	21
Purdue	3.44	3.49	3.03	32
Cal Tech	3.39	4.07	3.88	8
Duke	3.38	3.38	3.63	14

Table 4: A simple bibliometric rating of departments and its components vs. reputation rating