Data Mining Citations to Predict Emerging Scientific Leaders and Citation Curves

Peter Z. Revesz

***Abstract*—**Identifying emerging scientific leaders using the rapidly and continuously growing scientific citation databases is an important challenge. Any researcher's, yielding a citation curve. We describe a novel method for predicting many years ahead any researcher's citation curve (the ordering from the highest-cited to the lowest-cited of the researcher's publications). The method depends on treating the citation curves of researchers for various years as one single spatio-temporal function from rank and time to citations. For each researcher, we derive an estimate of this spatio-temporal function that can be used to predict the total citations of individual publications at any given rank at any time. This method can accurately predict entire citation curves and derived measures, such as, the total citations to all publications and the h-index of the researchers.

***Keywords*—**citation, forecasting, h-index, prediction, Web of Science.

I. Introduction

THE number of scientific journal publications is over 1.5 million per year and growing exponentially according to the US National Science Foundation's publication *Scientific and Engineering Indicators-2014*. Since each scientific journal article is a complex text that can be indexed and searched using various keywords and authors, the scientific literature is a big data that requires novel and efficient algorithmic methods to handle. On any single topic, it is usually impossible to recall all publications that may be relevant to a search topic. Hence a common approach to identify the most important scientific publications and authors is to measure their citations. High citations to a researcher's scientific publications are commonly considered a sign of accomplishment and prominence in his or her field of research. For example, Garfield [4] and Gingras and Wallace [5] found that until the 1960s a significant percentage of the Nobel Prize winners came from the top 500 highest cited authors. Therefore, much data has been compiled to provide citation statistics for individual researchers, for example, the *Web of Science* database by Thomson Reuters and the online *Google Scholar* database by Google. Using these databases, one can identify prominent researchers. For example, each year Thomson Reuters identifies the top one percentile of the highest cited researchers in several different fields of research and provides their names in a list of *Highly Cited Researchers*.

Peter Z. Revesz is a professor in the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588 USA (phone: 402-472-3488; fax: 402-472-7767; email: revesz@cse.unl.edu). The initial phase of this research was presented at a conference [15] and was supported by the *US Air Force Office of Scientific Research* where Dr. Revesz served as a Program Manager from 2012-14 on a leave from the University of Nebraska-Lincoln.

However, the problem of citation databases is that researchers are normally past middle age by the time they can be identified as citation leaders in their fields. The same problem plagues other measures of excellence that are based on citations, such as, the h-index measure proposed by J.E. Hirsch [6].

In contrast, in many situations people, such as, members of scientific hiring and promotion committees and grant review panels, are charged with evaluating the research potential of candidates. In that context, it may be beneficial for these people to predict the future citations of candidates as one measure of research potential. Hence we propose a data mining method that uses past citations data to predict future citations. In fact, we predict for any time t in the future and for any individual researcher his or her citation curve, which obtained at any time by ordering from highest to lowest-cited the publications of the researcher. For example, Andre Geim's scientific publications at the end of 2012 yield the citation curve shown in Fig. 1.

Predicting entire citation curves is a novel task. Previous authors, like Acuna [1] predicted the h-index, which is only a single point of the citation curve, or like Ponomarev et al. [9] and Wang et al. [16] proposed methods for identifying emerging scientific publications, i.e., predicting the long-term impact of single scientific publications. The latter authors commonly look for "breakthrough," "revolutionary," "game changing" or "seminal" publications and start their search by preselecting publications that have a certain minimum number of citations. That number tends to be large and overlook many highly cited publications. Hence it does not give a clear picture of most individual researchers' potential.

For example, Andre Geim and Konstantin Novoselov, the 2010 Nobel Prize winners in Physics for their work on graphine have a single publication that has received 9,152 citations by the end of 2012 as listed in the Web of Science database. In fact, Andre Geim had a total of 48,414 citations at that time. However, his total citations was surpassed by Mildred Dresselhaus, a leading carbon chemistry researcher at MIT, and one who is often considered a candidate for the Nobel Prize. At the end of 2012, she had 741 publications, which received a total of 49,102 citations according to the Web of Science database. An approach looking for only breakthrough research may identify Geim and Novosolev's graphine paper, but it may overlook Dresselhaus' publications. Hence approaches that look for only breakthroughs do not necessarily identify all future scientific leaders, which clearly all three authors are.

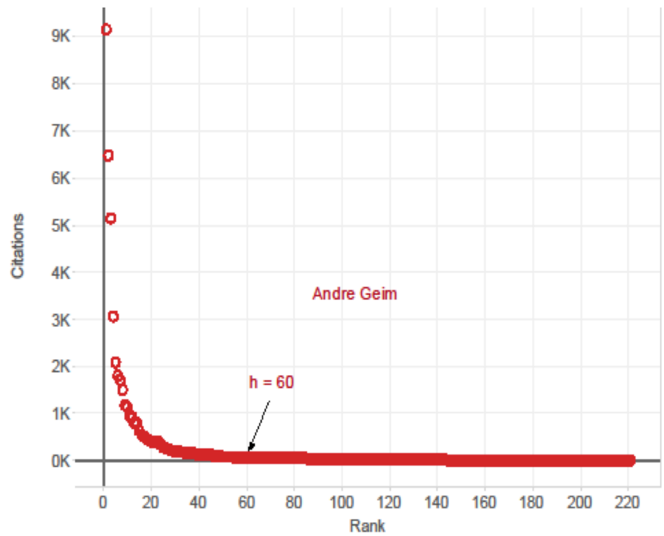

Fig. 1 The citation curve of Andre Geim, the 2010 Nobel Prize winner in physics. At the end of 2012, he had 221 publications with a total of 48,414 citations as listed in the Web of Science publications database.

A naïve approach to predict the citation curve of an individual researcher A would be to identify the future citations of each publication of A. There are two problems with that approach.

1. First, predicting the future citations of individual publications is very difficult, and many previous attempts were unsuccessful. Even a highly cited publication could be superseded by another publication and lose favor among future authors. Some contributions of other publications become so well-known that many textbooks carry detailed discussions of them, and in this case many authors prefer to cite the textbooks instead of the original publications.
2. Second, even if we could predict perfectly at time t the future citations of the publications that were written by that time by researcher A, we still would not know how many other publications the person will write. Therefore, we cannot predict future citation curves simply by predicting the future of already existing publications.

In contrast, our approach, whose earlier version was presented in [15], is to consider the changing and expanding citation curves as moving or spatio-temporal objects expanding our research in that area [2, 3, 7, 8, 10, 11, 12, 13, 14]. We develop an approximation of the spatio-temporal citation curve by a mathematical function from rank and time to citations, where in the approximation the domain of rank and time and the range of citations is the set of rational numbers. Such a generalized spatio-temporal approximation is derived from spatial approximations of the citation curves at a fixed number of already past time instances. In turn, these spatial approximations are complex in themselves because they require several pieces each. Experimentally we found good approximations with only three pieces for most researchers' citation curves, namely, one piece for the top forty percentile of publications, a second piece for the middle twenty percentile of publications, and a third piece for the bottom forty percentile of publications.

The spatio-temporal citation curve allows predicting at any future time many measures that are based on citation curves, such as, total citations of each publication and the h-index, which is the largest rank where the citations are still greater than equal to the rank [6]. Indeed, our spatio-temporal approximation can be used to analyse citation curves and answer many complex queries, such as, "What will be the h-index of researcher A at time t?" or "When will the h-index of researcher A exceed the h-index of researcher B?" or "When will researcher C have at least ten papers that are each cited over 100 times?

This paper is organized as follows. Section II gives some basic definitions. Section III describes our method of approximating citation curves at fixed time instances. Section IV describes our prediction method. Section V gives experimental results. Finally, Section VI presents some conclusions and future work.

II. Basic Concepts

Let the publications of researcher A be A_i for $1 \leq i \leq n$ for n publications in order of citation rank from highest to lowest cited. Such an ordering yields a citation curve from rank to citations, which are both integers. We denote by $c(A_i, t)$ the cumulative total citation count of publication A_i at time t, which is usually the end of some year. We denote by c(A, t) the total citation count for all publications of A_i for $1 \leq i \leq n$ at time t. Similarly, we denote by e(Ai, t) the estimated cumulative total citation count of publication A_i at time t. We denote by e(A, t) the estimated total citation count for all publications A_i for $1 \leq i \leq n$ at time t. Finally, we denote by h(A, t) the estimated h-index of researcher A at time t. An h-index of researcher A is the number of publications that has greater than equal citations that their ranks.

In this paper, we propose several estimation methods to find $e(A_i, t)$, which gives an estimate of the entire citation curve of researcher A at time t. We are primarily interested in finding estimates with low mean squared error (MSE), defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (e(A_i, t) - c(A_i, t))^2$$

The square root of the MSE is the standard error. For the computer experiments in Section 5, the standard error, or StdErr, is calculated using the Tableau data analytics and visualization system. Regarding e(A, t) and h(A, t), we are interested in minimizing the percentage difference from c(A, t) and the real h-index value at time t.

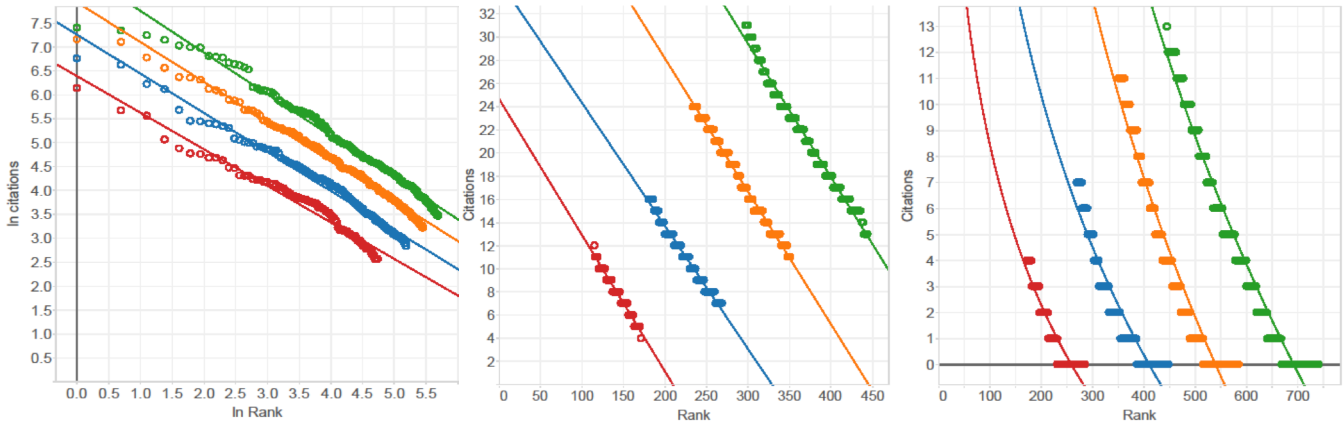

Fig. 2 Dresselhaus' citation curves at the end of 2000 (red), 2004 (blue), 2008 (orange) and 2012 (green) are shown for the top 40 percentile of the publications (left), for the 40 to 60 percentile of publications (middle) and for the bottom 40 percentile of publications (right). Note that the graph on the left is displayed as a natural logarithm of rank (x axis) and natural logarithm of citations (y axis).

III. Citation Curves Approximated by Functions

A researcher's citation curve is piecewise approximated by functions of the form:

$$y = a\,x^{p} + b \qquad (1)$$

where x is the rank, y is the total citations of the publication of that rank, and a, b and p are rational number constants. While the original citation curve has a discrete domain, the approximation power law function f(x) has a rational domain. We approximate each researcher's citation curve in three pieces using equation (1) with the following restrictions:

Top 40 percentile of publications: b = 0, that is, a power law function. The top 40 percentile of Dresselhaus' citation curves in years 2000, 2004 and 2008 can be visualized on a log-log graph as shown in Figure 2A. The near linearity of the curves shows that there is a power law relationship between rank and citations.

40 to 60 percentile of publications: p = 1, that is, a linear function. The linearity of these citation curve pieces is shown in Figure 2B.

Bottom 40 percentile of publications: a = 0, that is, a constant function. Moreover, b = ln(Min60), i.e., the natural logarithm of Min60, which is the total citations to the last publication in the top 60 percentile. Figure 2C shows that for the bottom 40 percentile of publications the citations vary with the logarithm of the rank. We found ln(Min60) to be a good approximation.

Example 1. For Dresselhaus' publications at the end of 2000, 2004, 2008 and 2012, we found using the Tableau data analytics and visualization system the best-fit approximations shown in Table 1. The functions in Table 1 give close approximations to Dresselhaus' citation curves. Table 2 shows the approximate and the actual citations of some publications of Dresselhaus between 2000 and 2012.

Table 1 Approximations for Dresselhaus' citation curves in 2000, 2004, 2008 and 2012

Year	Percent	a	p	b	Min_{60}
2000	0-40	595	0.7645	0	
2004	0-40	1418	0.8176	0	
2008	0-40	2785	0.8318	0	
2012	0-40	5568	0.8715	0	
2000	40-60	0.1177	1	24.674	
2004	40-60	0.1062	1	34.941	
2008	40-60	0.1147	1	51.103	
2012	40-60	0.1154	1	64.078	
2000	60-100				4
2004	60-100				7
2008	60-100				11
2012	60-100				13

Table 2 Tests of the approximations on some sample publications

Year	Percentile	Rank	Citations Approximate	Citations Actual
2000	0-40	75	21.94	21
2004	0-40	75	41.57	45
2008	0-40	75	76.78	83
2012	0-40	75	129.28	133
2000	40-60	150	7.02	7
2004	40-60	200	13.70	13
2008	40-60	250	22.43	23
2012	40-60	400	17.91	18

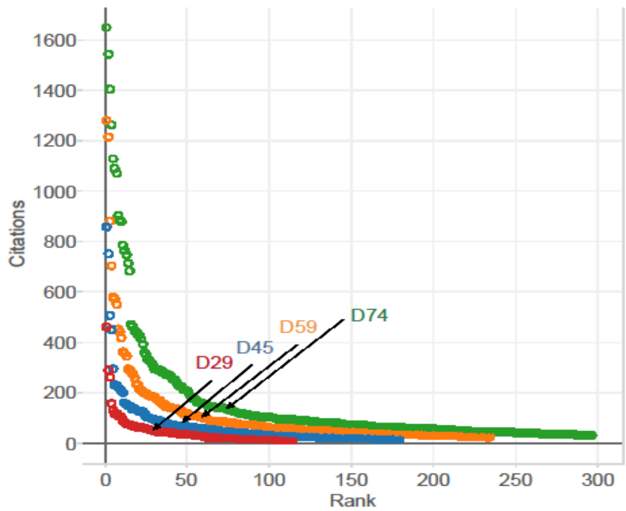

Fig. 3 The top 40 percentile of Dresselhaus' citation curves at the end of 2000 (red), 2004 (blue), 2008 (orange) and 2012 (green). On each of the curves the publications closest to the 10th percentile, that is, the arrows mark D29 in 2000, D45 in 2004, D59 in 2008 and D74 in 2012.

IV. A Citation Prediction Method

Our citation prediction method is based on extending the approximate citation curves $f_0(x)$, $f_1(x)$, $f_2(x)$ ... at fixed times t_0, t_1, t_2 ... respectively, into a spatio-temporal citation curve $\varphi(x, t)$. Figure 3 shows the top 40 percentile of Dresselhaus' citation curves at the end of 2000 (red), at the end of 2004 (blue), at the end of 2008 (orange) and at the end of 2012 (green). These four separate citation curves can be viewed as instances of a spatio-temporal citation curve. Intuitively, in the spatio-temporal citation curve the entire curve expands proportionally. Hence publications at the same percentile correspond to each other. For example, the publications nearest the 10^{th} percentile in each citation curve, that is, publications D_{29} in 2000, D_{45} in 2004, D_{59} in 2008 and D_{74} in 2012 are placed at similar locations on the corresponding citation curves. That correspondence is similar to the correspondence of items that are each one standard deviation above the mean in two normal distribution graphs, one distribution generated from a small sample and the other distribution generated from a large sample of the same population.

The citation prediction method can be described in a theorem as follows.

Theorem 1 For any researcher, let $f_0(x)$, $f_1(x)$ and $f_2(x)$ be the approximation functions for the citation curves and n_0, n_1 and n_2 be the number of publications at times t_0, t_1 and t_2, respectively, where $t_2 > t_1 > t_0$. Then the researcher's spatio-temporal citation curve $\varphi(x, t)$ for any time $t > t_2$ can be estimated to be:

$$\begin{aligned}\varphi(\mathrm{x},\mathrm{t}) = &\left(\frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) f_0\left(\frac{\mathrm{n}_0\mathrm{x}}{n(t)}\right) \\ &- \left(\frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2} + \frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) f_1\left(\frac{\mathrm{n}_1\mathrm{x}}{n(t)}\right) \\ &+ \left(1 + \frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2}\right) f_2\left(\frac{n_2 x}{n(t)}\right) \qquad (2)\end{aligned}$$

where $n(t)$ is the number of publications at time t. In addition, $n(t)$ could be estimated by:

$$\begin{aligned}\mathrm{n}(\mathrm{t}) = &\left(\frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) n_0 \\ &- \left(\frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2} + \frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) n_1 \\ &+ \left(1 + \frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2}\right) n_2 \qquad (3)\end{aligned}$$

Proof: By the proportionality assumption, rank x at time t corresponds to rank n_0*x/n, n_1*x/n and n_2*x/n at times t_0, t_1 and t_2, respectively. For simplicity, let $C_0 = f_0\left(\frac{\mathrm{n}_0\,\mathrm{x}}{n(t)}\right)$, $C_1 = f_1\left(\frac{\mathrm{n}_1\,\mathrm{x}}{n(t)}\right)$ and $C_2 = f_2\left(\frac{\mathrm{n}_2\,\mathrm{x}}{n(t)}\right)$. Between corresponding ranks, the velocity of the citations change at t_1 is:

$$\mathrm{V}_1 = \frac{\mathrm{C}_1 - \mathrm{C}_0}{\mathrm{t}_1 - \mathrm{t}_0}$$

Similarly, the velocity of the citations change at t_2 is:

$$\mathrm{V}_2 = \frac{\mathrm{C}_2 - \mathrm{C}_1}{\mathrm{t}_2 - \mathrm{t}_1}$$

Simplifying yields:

$$\begin{aligned}\varphi(\mathrm{x},\mathrm{t}) = &\left(\frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) C_0 \\ &- \left(\frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2} + \frac{(t-t_2)^2}{2\,(t_2-t_1)(t_1-t_0)}\right) C_1 \\ &+ \left(1 + \frac{t-t_2}{t_2-t_1} + \frac{(t-t_2)^2}{2\,(t_2-t_1)^2}\right) C_2\end{aligned}$$

The above yields Equation (2), and the approximation for $n(t)$ can be obtained similarly. QED

Example 2. Table 1 implies that the top 40 percentile of Dresselhaus' citation curves for 2000, 2004 and 2008 can be approximated by the following power law functions:

$f_{2000,\,0\text{-}40} = 595.368\ x^{-0.7645}$
$f_{2004,\,0\text{-}40} = 1418.237\ x^{-0.8176}$
$f_{2008,\,0\text{-}40} = 2785.214\ x^{-0.8318}$

We also know that $n_{2000} = 286$, $n_{2004} = 449$ and $n_{2008} = 585$. Substituting these into Equation (2) yields:

$$\varphi_{0-40}(x,t) = 297.68\left(\frac{286\,x}{n(t)}\right)^{-0.7645} - 2836.47\left(\frac{449\,x}{n(t)}\right)^{-0.8176} + 6963.04\left(\frac{585\,x}{n(t)}\right)^{-0.8318}$$

Similarly, Table 1 implies for the 40-60 percentile of Dresselhaus' citation curves the approximations:

$f_{2000,\,40\text{-}60} = -0.117708\,x + 24.674$
$f_{2004,\,40\text{-}60} = -0.106194\,x + 34.941$
$f_{2008,\,40\text{-}60} = -0.114707\,x + 51.103$

Substituting these into Equation (2) yields:

$$\varphi_{40-60}(x,t) = -\ 0.058854\left(\frac{286\,x}{n(t)}\right) + 0.212388\left(\frac{449\,x}{n(t)}\right) - 0.2867675\left(\frac{585\,x}{n(t)}\right) + 70.2127$$

In particular, at exactly the 60 percentile of the approximation graph, we have for any t:

$$\mathrm{Min}_{60} = -\,0.058854\,(286 * 0.6) + 0.212388\,(449 * 0.6) - 0.2867675\,(585 * 0.6) + 70.2127 \approx 16.7$$

Hence, we estimate:

$$\varphi_{60-100}(x,t) = \ln(16.7)$$

The functions $\varphi_{0-40}(x,t)$, $\varphi_{40-60}(x,t)$ and $\varphi_{60-100}(x,t)$ form the piecewise approximation $\varphi(x,t)$ of the citation curve.

We can analyse the spatio-temporal citation curve of any researcher at any time the same way as the actual citation curves are analysed. For example, we can easily find an estimate of the total citations or the h-index.

Theorem 2 Let any researcher A's spatio-temporal citation curve be *φ(x, t)* and temporal publication function be *n(t)*. Then A's total citations and h-index at time t can be estimated as follows:

$$c(A,t) = \sum_{i=1}^{n(t)} \varphi(i,t) \qquad (4)$$

$$h(A,t) = i \quad \text{if } \varphi(i,t) \geq i \text{ and } \varphi(i+1,t) < i+1 \qquad (5)$$

Example 3. Suppose that we want to find the total citations and the h-index for Dresselhaus in 2012 using data from 2000, 2004 and 2008. Substituting $t_0 = 2000$, $t_1 = 2004$, $t_2 = 2008$, and $t = 2012$, $n_{2000} = 286$, $n_{2004} = 449$ and $n_{2008} = 585$ into Equation (3), we estimate the number of publications in 2012 as:

$$n_{2012} = 2.5\ n_{2008} - 2\ n_{2004} + 0.5\ n_{2000} \approx 708$$

Hence we also assume that the top 40 percentile of publications run from D_1 to D_{283}, the middle twenty percentile from D_{284} to D_{425}, and the bottom twenty percentile from D_{426} to D_{708}. Hence substituting into Equation (4) we estimate of the total citations of the publications in the top 40 percentile in 2012 as follows:

$$C_{0-40}(A,2012) = \sum_{i=1}^{283} 297.68\left(\frac{286\,i}{708}\right)^{-0.764526} - 2836.47\left(\frac{449\,i}{708}\right)^{-0.817568} + 6963.04\left(\frac{585\,i}{708}\right)^{-0.831757}$$

The above estimate is about 45,929. The actual number of total citations for the top 40 percentile of publications was 44,617 in 2012. For the middle 20 percentile in 2012, we estimate:

$$C_{40-60}(A,2012) = \sum_{i=284}^{425} -\,0.058854\left(\frac{286\,i}{708}\right) + 0.212388\left(\frac{449\,i}{708}\right) - 0.2867675\left(\frac{585\,i}{708}\right) + 70.2127$$

The above estimate is 3,626. We note that $\mathrm{Min}_{60} = 16.7$. Hence for the bottom 40 percentile we estimate:

$$C_{60-100}(A,2012) = \sum_{i=426}^{708} \ln(16.7) \approx 796$$

Hence our citation prediction method estimates that c(A, 2012) = 45,929 + 3,626 + 796 = 50,351 citations. That estimate was within 2.5% of the actual value.

We can use Equation (5) to estimate the h-index too in 2012 by simply calculating in a while loop the estimated citations for each paper until the rank exceeds the total citations for a publication. Figure 4 shows a comparison between the actual and the estimated citation curves of Dresselhaus for 2012. As can be seen the two curves are close to each other. Hence the actual and the estimated h-index values, 101 and 99, respectively, are also close to each other.

V. Experimental Results

A. Sample Data Collection and its Statistical Analysis

In our experiments we collected citation data from the Web of Science database for eight leading physics researchers. These researchers included Andre Geim and Konstantin Novosolev, winners of the Nobel Prize in Physics in 2010, Serge Haroche and David Wineland, winners of the Nobel Prize in Physics in 2012, Brian P. Schmidt, one of the winners

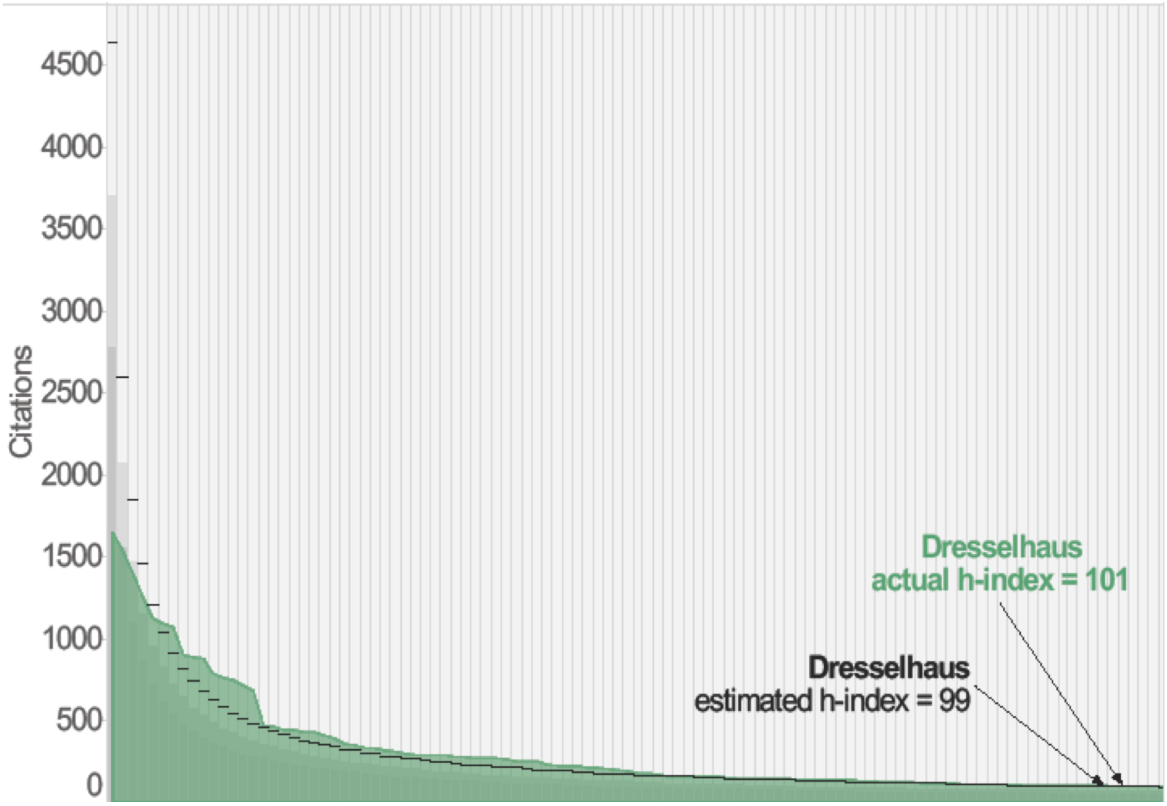

Fig. 4 Dresselhaus' actual 2012 citation curve (green area) and estimated 2012 citation curve (black line) for publications ranked from 1 to 105. The estimate was based on citations data from 2000, 2004 and 2008. The close relationship between the two curves allows good estimation of other measures that are based on citations, such as, the h-index. In this case, the actual h-index is 101 and its estimate is 99.

of the Nobel Prize in Physics in 2011, and Mildred Dresselhaus, Geoffrey Marcy, and Didier Queloz. The last three of these researchers are also rumored to be nominees for the Nobel Prize. All of these prominent researchers have a large set of publications, which enabled easier data collection from the *Web of Science* database and statistically more reliable experiments.

Our first goal was to test the distribution of the citations of these researchers. Fig. 5 shows that their citations are approximately log normal distributed. Since other researchers already found lognormal distributions for the citations of large number of researchers, the lognormal distribution of our data set suggests that we have a good representative sample in terms of statistical distribution even though even each of the selected researchers had a larger number of publications than usual. Previously, one of our concerns was that the Web of Science database may ignore many low impact journals, which would result in a lower than expected number of non-cited publications. In contrast, the distribution in Fig. 5 deviated from a perfect lognormal distribution in a somewhat larger number of non-cited or rarely cited publications. That deviation which was probably due to the fact that all of the researchers were still active, that is, they continued to publish many new publications that had little chance yet to be read and cited by other researchers.

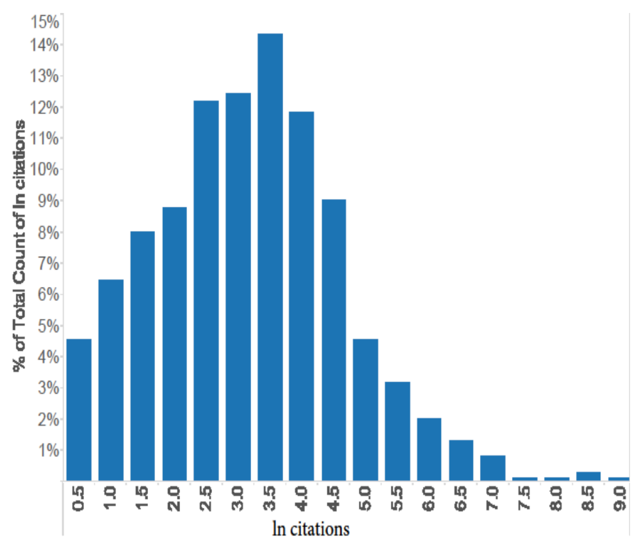

Fig. 5 The distribution of the natural logarithms of the citations of the publications of eight physics researchers is approximately a normal distribution. The larger than expected percentage of publications with very low natural logarithm of citations is due mainly to recently published publications that have not received any citations because of their recentness. A more symmetric distribution can be obtained if we only consider publications that were published at least five years ago.

Table 3 The experimental results for the citation curves for 2012.

ID	Percent	a	p	b	Min_{60}	Sum Estimate	Sum Actual
Dresselhaus	0-40	5568.383	-0.871541	0		49871	44617
Dresselhaus	40-60	-0.115414	1	64.0783		3151	3155
Dresselhaus	60-100				13	762	1330
Dresselhaus	0-100					53784	49102
Geim	0-40	25714.235	-1.44745	0		65392	47198
Geim	40-60	-0.54179	1	80.8223		922	922
Geim	60-100				12	221	294
Geim	0-100					66535	48414
Haroche	0-40	1938.675	-0.8687	0		10492	9917
Haroche	40-60	-1.65895	1	125.39		806	806
Haroche	60-100				22	133	309
Haroche	0-100					11431	11032
Marcy	0-40	968.7	-0.621212	0		13276	12750
Marcy	40-60	-0.321623	1	74.0623		1617	1616
Marcy	60-100				20	339	765
Marcy	0-100					15232	15131
Novoselov	0-40	28787.6	-1.56564	0		63787	42834
Novoselov	40-60	-0.967925	1	96.1688		615	615
Novoselov	60-100				6	115	101
Novoselov	0-100					64517	43550
Queloz	0-40	1013.292	-0.707164	0		10899	10931
Queloz	40-60	-0.250972	1	53.1941		1015	1015
Queloz	60-100				12	281	387
Queloz	0-100					12195	12333
Schmidt	0-40	2502.486	-0.935619	0		14797	17519
Schmidt	40-60	--0.39078	1	69.1014		1042	1042
Schmidt	60-100				12	246	327
Schmidt	0-100					16085	18888
Wineland	0-40	2667.268	-0.835968	0		17536	16185
Wineland	40-60	-0.826738	1	107.235		1327	1327
Wineland	60-100				25	212	438
Wineland	0-100					19075	17950

B. *The Accuracy of the Approximation for Citation Curves*

Next we mapped all the researchers' top 40 percentile citation curves onto a log-log graph, except for Dresselhaus, whose citation curve we displayed already in Figures 2 and 3. The left side of Figure 6 shows that for the seven other seven researchers also the top 40 percentile tended to have a power law relationship between rank and citations of publications at the end of 2012. The right side of Figure 6 shows a linear relationship for the middle twenty percentile. Hence the trends noticed for Dresselhaus' publications hold as well for the other researchers.

Table 3 shows the details of the piecewise linear approximations. In all cases the standard error was below 0.0736611, and the p-value was below 0.0001. The difference between the estimated and the actual total citations were less than 15% for all researchers except Geim and Novoselov, who authored on graphine a celebrated and highly cited publication, which accounts for most of the differences between their estimated and actual total citations.

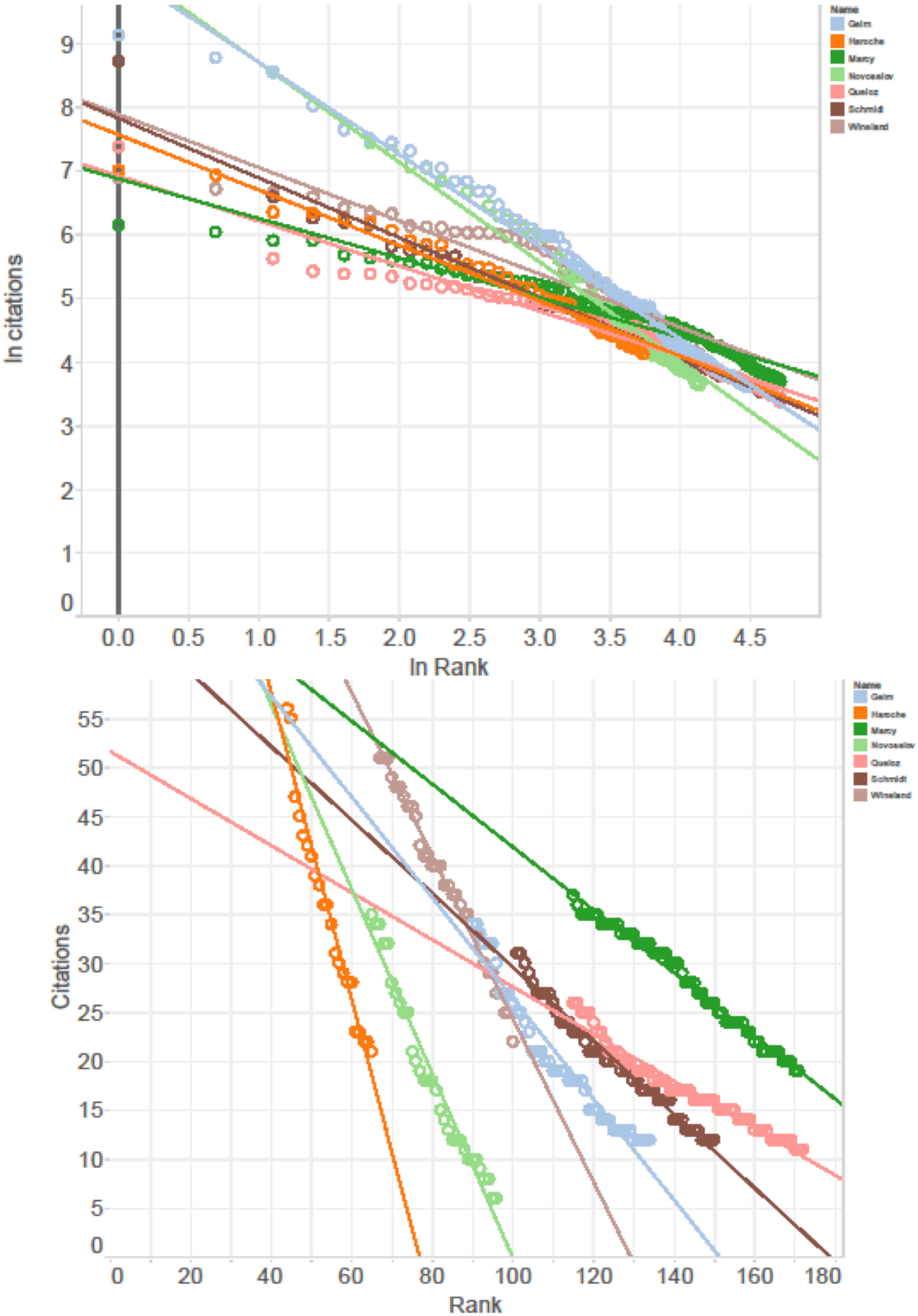

Fig. 6 A power law between rank and citations describes the top 40 percentile of publications of the seven other leading physics researchers in 2012 (left). A linear function between rank and citations describes the 40-60 percentile of publications of the seven other leading physics researchers in 2012 (right).

C. Experiments Testing the Accuracy of Predicting Total Citations and H-index

We tested the accuracy of predicting 2012 total citations and h-index using data from 2000, 2004 and 2008. The results for Wineland are shown in Table 4 where the 2012 estimate was within 0.7 percent of the actual for total citations and 1.6 for the h-index.

Table 4 The experimental results for Wineland: The piecewise approximations for 2000, 2004 and 2008 and the predictions for total citations and h-index for 2012 (in last two rows).

Year	Percent	a	p	b	Min60	Sum Est.	Sum Act.
2000	0-40	530.28	-0.760	0		3105	3029
2000	40-60	-1.0786	1	62.85		328	328
2000	60-100				15	84	104
2000	0-100					3517	3461
2004	0-40	1130.45	-0.799	0		6899	6585
2004	40-60	-1.2247	1	94.64		650	650
2004	60-100				18	121	208
2004	0-100					7670	7443
2008	0-40	1693.83	-0.787	0		11401	10795
2008	40-60	-1.17	1	111.4		912	912
2008	60-100				24	165	381
2008	0-100					12478	12088
2012	0-100					17833	17950
2012						60	61

VI. Conclusions and Future Work

We gave a prediction method for citations to all the publications of an individual researcher. We also experimented with a small set of physics researchers. The experiments show that the citation prediction method gave estimates that were close to the actual data. The experiments focused on Nobel Prize winners because their have extensive and well-documented publication records. The method's good performance on the early career years of these celebrated scientists, that is, when they had only a modest number of publications and citations, suggests that the method could work also well for ordinary researchers.

In the future, we would like to extend the experiments to a larger set of researchers and to include researchers in other scientific areas, including computer science, biology and chemistry. We also plan to experiment with approximations that are more refined, i.e., consist of more than just three pieces. Such approximations could further enhance the accuracy of the prediction method. We would also like to make experiments that test how accuracy changes with the number of years by which we try to predict ahead the total citations and the h-index.

References

[1] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success," *Nature*, 489 (7415), 201-202, September 2012.

[2] S. Anderson and P. Z. Revesz, "Efficient MaxCount and threshold operators of moving objects," *Geoinformatica*, 13 (4), 355-396, 2009. Available at: http://dx.doi.org/10.1007/s10707-008-0050-7

[3] J. Chomicki and P. Z. Revesz, "Constraint-based interoperability of spatiotemporal databases," *Geoinformatica*, 3 (3), 211-243, September 1999. Available at: http://dx.doi.org/10.1023/A:1009849314891

[4] E. Garfield, "Do Nobel Prize winners write citation classics?" *Essays of an Information Scientist,* 9, 182-187, 1986.

[5] Y. Gingras and M. L. Wallace, "Why it has become more difficult to predict Nobel Prize winners: A bibliometric analysis of nominees and winners of the chemistry and physics prizes (1901-2007)," *Scientometrics* 82 (2), 401-412. 2010.

[6] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Natl. Acad. Sci. USA,* 102, 16569-16572, 2005.

[7] P. C. Kanellakis, G. M. Kuper, and P. Z. Revesz, "Constraint query languages," *Journal of Computer and System Sciences*, 51 (1), 26-52, 1995. Available at: http://dx.doi.org/10.1006/jcss.1995.1051

[8] L. Li and P. Z. Revesz, "Interpolation methods for spatio-temporal geographic data," *Computers, Environment and Urban Systems*, 28 (3), 201-227, 2004. Available at: http://dx.doi.org/10.1016/S0198-9715(03)00018-8

[9] I. V. Ponomarev, D. E. Williams, C. J. Hackett, J. D. Schnell and L. L. Haak, "Predicting highly cited publications: A method for early detection of candidate breakthroughs," *Technological Forecasting and Social Change*, 81, 49-51, 2014.

[10] P. Z. Revesz, P. Z, *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, New York, NY, 2010.

[11] P. Z. Revesz, "On the semantics of arbitration," *International Journal of Algebra and Computation*, 7 (2), 133-160, 1997. Available at: http://dx.doi.org/10.1142/S0218196797000095

[12] P. Z. Revesz and T. Triplet, "Temporal data classification using linear classifiers," *Information Systems*, 36 (1), 30–41, 2011. Available at: http://dx.doi.org/10.1016/j.is.2010.06.006

[13] P. Z. Revesz and T. Triplet, "Classification integration and reclassification using constraint databases," *Artificial Intelligence in Medicine,* 49 (2), 79-91, 2010. Available at: http://dx.doi.org/10.1016/j.artmed.2010.02.003

[14] P. Z. Revesz and S. Wu, "Spatiotemporal reasoning about epidemiological data," *Artificial Intelligence in Medicine*, 38 (2), 157-170, 2006. Available at: http://dx.doi.org/10.1016/j.artmed.2006.05.001

[15] P. Z. Revesz, "A method for predicting citations to the scientific publications of individual researchers," *Proc. International Conference on Database Engineering and Applications Symposium (IDEAS'14),* Porto, Portugal, July 2014.

[16] D. Wang, C. Song and A. L. Barabási, Quantifying long-term scientific impact. *Science* 342 (6154), 127-132, 2013. Available at: http://dx.doi.org/10.1126/science.1237825

Peter Z. Revesz (Ph.D'91) holds a Ph.D. degree in Computer Science from Brown University and was a postdoctoral fellow at the University of Toronto.

He is an expert in databases, data mining, big data analytics and bioinformatics. He is the author of *Introduction to Databases: From Biological to Spatio-Temporal* (Springer, 2010) and *Introduction to Constraint Databases* (Springer, 2002). He is currently a professor in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln, Lincoln, NE 6815, USA.

Dr. Revesz also held visiting appointments at the Aquincum Institute of Technology, the IBM T. J. Watson Research Center, INRIA, the Max Planck Institute for Computer Science, the University of Athens, the University of Hasselt, the U.S. Air Force Office of Scientific Research and the U.S. Department of State. He is a recipient of an AAAS Science & Technology Policy Fellowship, a J. William Fulbright Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award, and a "Faculty International Scholar of the Year" award by *Phi Beta Delta*, the Honor Society for International Scholars.