# Predictive Ranking of Computer Scientists Using CiteSeer Data

Dror G. Feitelson Uri Yovel
School of Computer Science and Engineering
The Hebrew University of Jerusalem
91904 Jerusalem, Israel
Email: feit@cs.huji.ac.il

**Abstract**

The increasing availability of digital libraries with cross-citation data on the Internet enables new studies in bibliometrics. We focus on the list of 10,000 top-cited authors in computer science available as part of CiteSeer. Using data from several consecutive lists, we construct a model of how authors accrue citations with time. By comparing the rate at which individual authors accrue citations with the average rate, we make predictions of how their ranking in the list will change in the future.

**Keywords:** prediction, citation analysis, ranking, digital library

*It's tough to make predictions, especially about the future.*

– *Yogi Berra, American Baseball Player*

## 1 Introduction

While it is widely recognized that scientists are not all of the same caliber, it is not easy to measure this effect directly. As a result, indirect metrics have been proposed. One of the most common is counting citations (Garfield, 1979). The idea is that if a given individual publishes significant groundbreaking research, this will be cited by others. The more citations, the larger the impact. Counts of citations have thus been used for the ranking of individuals, universities, and even nations (May, 1997; ScienceWatch). They are also used to rank publication venues in the Journal Citation Reports.

Since the early 1960s, citation data has been available from the Science Citation Index. Creating this index has traditionally been a laborious process (Garfield, 1979). But with the advent of the Internet and digital libraries, a lot of data has become directly available in computer-accessible formats. CiteSeer is one of the major sites that exploits this opportunity, and provides a wealth of data for free. We build on this data, and provide an additional level of analysis.

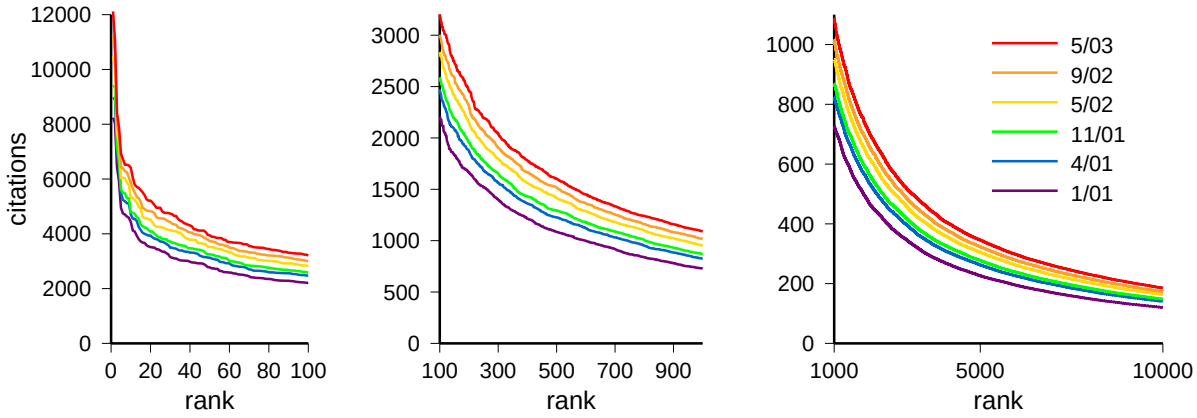


Figure 1: *Ranking and citation-count data from CiteSeer. Due to the large scale of citations, the list is divided into three regions.*

So far, the study of bibliometrics has been largely descriptive. But it is also possible to follow trends and make predictions rather than just report snapshots. One example was provided by Holmes and Oppenheim, who attempted to predict the results of the RAE evaluation in the UK based on current citation data (Holmes and Oppenheim, 2001). Another was provided by Geller at al., who attempt to predict the total number of citations that papers will accrue over a lifetime of 40 years (Geller et al., 1981). Our work is, in a sense, a combination of the two: we predict ranking based on predictions of citations.

# 2 The CiteSeer Database

CiteSeer is a web site providing reference data in computer science (CiteSeer). It is based on autonomous citation indexing (Lawrence et al., 1999). The idea is to perform an automated search of the Internet, starting with home pages of universities and research centers. The goal of the search is to harvest research papers that have been posted on personal home pages. These are parsed to extract the authors, titles, and citations. The data is then tabulated and analyzed, enabling the site to rank papers according to how many citations they have received.

It should be noted from the outset that the data provided by CiteSeer is far from perfect. Not all research papers are posted on the Internet, and in particular, papers actually published in journals may be missing due to copyright limitations. Others may be posted several times in different versions, leading to double counting. Parsing various citation formats to extract citation information is also problematic (but this has improved considerably over the years). However, the site has earned a wide following and is considered a very important and useful resource for searching the computer science literature. It contains data about hundreds of thousands of papers, and tends to be more up-to-date than traditional citation services (Goodrum et al., 2001).

In addition to enabling a keyword- and author-based search of the database, CiteSeer also

publishes compiled statistics. One of the most popular is the list ranking the 10,000 most cited authors in computer science (out of a total of 659,481 and growing in the database). This list is updated at irregular intervals of several months (Fig. 1). Each list provides a snapshot that captures the ranking for the time of its publication. But naturally, ranks change over time. For example, it is plausible that the people who will be the top ranking researchers in 30 years are now only at the beginning of their careers, and do not have so many citations yet. Our goal is to try to predict the rank that a person will eventually achieve, based on how his ranking improves from one list to the next.

CiteSeer itself also provides predictions of future citation counts. The model used for predictions is extremely simple: the citation count to each paper is multiplied by a factor that depends on the number of years since it was published. This is a simplification of the model of (Geller et al., 1981). The idea is that recent papers have not had sufficient time to accumulate all their citations yet, so the current count should be multiplied by a higher factor. This model has two severe limitations. First, the multiplicative factor is the same for all researchers in the database. It thus does not reflect possible differences between papers, researchers, or whole fields. Second, the predictions are based on a static snapshot of each person's publications. There is no consideration of possible future work, and citations it will receive (indeed, Geller et al. focus on ranking based on citations per paper, not on total citations; we prefer total citations as in CiteSeer).

Our model, in contradistinction, is based on modeling the evolution of citations to individual researchers. If a certain person accrues citations faster or slower than others, this will be included in the model. Moreover, the model implicitly includes the assumption that people continue to produce new papers that will also be cited at the same rate as their current papers. This is expected to be very meaningful for the ranking of prolific researchers who author many more papers than the average.

# 3 Theory of Predictive Ranking

A researcher's number of citations is the sum of citations for each paper he wrote. The number of papers grows with the length of the career. The number of citations accumulated by each paper grows with time, and may be expected to stabilize eventually, due to obsolescence. Ranking compares all this to similar processes occurring for other researchers at different stages of their careers.

Obviously, direct modeling would be very complex, and require correctly parameterized submodels for the paper production rates and the citation accumulation rates. A simpler approach is to treat the total accumulation of citations as an opaque process, rather than trying to model its mechanics. In addition, we focus on ranking rather than on the actual number of citations. This improves the reliability of the model, because ranking only depends on the relative number of citations, not on the absolute numbers. If the predicted numbers of citations turn out to be incorrect, but the deviations are consistent for all researchers, the ranking will still be reliable.

| | |
|---|---|
| $C(r,t)$ | citations at rank $r$ at time $t$ |
| $C_p(t)$ | citations of person $p$ at time $t$ |
| $R(c,t)$ | rank achieved with $c$ citations at time $t$ |
| $R_p(t)$ | rank of person $p$ at time $t$ |
| $R_p^\infty$ | asymptotic rank of person $p$ |

Table I: *Notation used in derivations.*

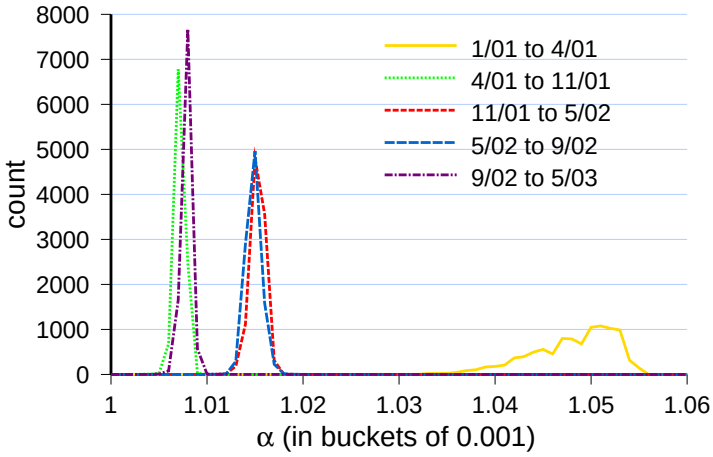


Figure 2: *The distribution of $\alpha$ for different pairs of ranking lists. $\alpha$ is the monthly growth rate in number of citations for a given rank.*

## 3.1 The Multiplicative Model

In the model, we consider the number of citations needed to achieve a certain rank in the CiteSeer list. By comparing the number of citations attributed to the same rank in consecutive lists, we find that a simple multiplicative model provides a very good match. We thus claim that (see Table I for notation)

$$C(r, t+1) = \alpha C(r,t)$$

Where $t$ is measured in months and $\alpha$ is a constant. Looking at longer time intervals, this leads to the prediction

$$C(r, t+\delta) = C(r,t)\alpha^\delta$$

Note that this relates to a rank, not to a certain person occupying that rank in a specific list. (Incidently, the multiplicative model implies that the distribution of citation counts should be lognormal, at least for the right tail of the distribution, which is the part we are looking at.)

The actual distribution of $\alpha$ for different pairs of lists is shown in Fig. 2. As can be seen, except for the first pair of lists the distribution is quite narrow, so the assumption that $\alpha$ is constant is reasonable. In addition, predictions that used data from the first list proved to be unreliable. We therefore decided to discard the first available list, and focus on the other five.

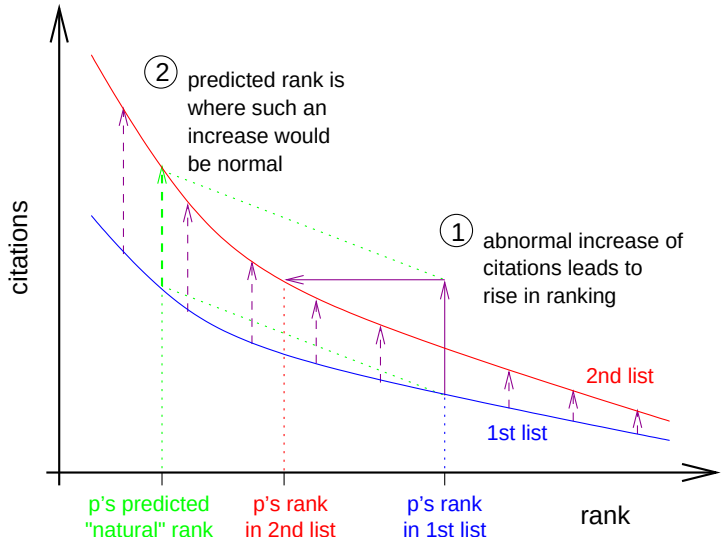


Figure 3: *Predicting the asymptotic ranking based on increase in citations.*

Regrettably, different pairs of lists lead to somewhat different values for $\alpha$. This may be caused by modifications to the methodology used by CiteSeer over time, e.g. improved coverage of the web or improved parsing of citations. The implication is that our ranking predictions will be based on pairs of lists, and use the correct $\alpha$ for each pair.

## 3.2 The Asymptotic Rank

The framework for predicting the future rank of a specific researcher $p$ is as follows. Assume we have two lists, from times $t-1$ and $t$. We assume that the citation counts at all ranks of the list grow at a monthly rate of $\alpha$. But the citation count of $p$ grows differently. Using data about $p$, we will attempt to predict his rank in consecutive months $t+1, t+2, \ldots$, and the asymptotic rank to which he will converge.

There is a simple consideration that allows us to predict the asymptotic value even without a detailed model. The idea is that $p$ will move up (or down) in the list until his rate of change of citations is equal to that of his neighbors. Assuming the lists are monotonic and smooth, and that the differences at the top are larger than at the bottom, we can simply find the rank where the difference matches $p$'s difference (Fig. 3). Define the difference in $p$'s citations

$$\Delta_p = C_p(t) - C_p(t-1)$$

This leads to the prediction

$$R_p^\infty = r \quad \text{s.t.} \quad C(r,t) - C(r,t-1) = \Delta_p$$

We call this the empirical method. Alternatively, we can find the matching rank by looking at only one list and using the definitions of $\alpha$. This gives

$$R_p^\infty = r \quad \text{s.t.} \quad (\alpha - 1)\, C(r,t-1) = \Delta_p$$

which we call the analytic method.

### 3.3 Dynamics

But what is the prediction for the immediate future? For this we need a model of how the number of citations changes, in comparison with that of other researchers. Focusing on researcher $p$, we can define $\alpha_p = C_p(t)/C_p(t-1)$. Assume that $p$ is a rising star, and therefore $\alpha_p > \alpha$. It would be wrong to predict $C_p(t+\delta) = C_p(t)\alpha_p^\delta$ because this will cause $p$'s number of citations to rise faster than those of anyone else, and on the long run he will necessarily become the top researcher. Likewise, if $\alpha_p < \alpha$, $p$ will necessarily lose ground and eventually will be dropped off the list. To achieve a situation in which $p$'s rank converges to a reasonable value, we need a functional form in which the dependence on time is $\alpha^\delta$. We therefore use only $\Delta_p$ as the part that grows with time:

$$\begin{aligned} C_p(t+\delta) &= C_p(t) + \Delta_p + \Delta_p\alpha + \Delta_p\alpha^2 + \cdots + \Delta_p\alpha^{\delta-1} \\ &= C_p(t) + \Delta_p\frac{1-\alpha^\delta}{1-\alpha} \end{aligned}$$

This provides a model of $p$'s number of citations. To turn this into a ranking, we need to find the number of citations $C$ at time $t$ that would have grown to this at the usual rate of $\alpha$. This satisfies $C\alpha^\delta = C_p(t+\delta)$, so $C = C_p(t+\delta)/\alpha^\delta$. The predicted rank of $p$ at time $t+\delta$ is then

$$\begin{aligned} R_p(t+\delta) &= R(C,t) \\ &= R\left(\frac{C_p(t) + \Delta_p\frac{1-\alpha^\delta}{1-\alpha}}{\alpha^\delta}, t\right) \end{aligned}$$

An open issue is how to handle newcomers, that only appear in the newer list. One option is to use the bottom number of citations from the old list as an upper bound on the number of citations the newcomers had at that time — had they had more, they would have been in the list. This translates to a lower bound on their rate of progress.

# 4 Reduction to Practice

The above model cannot be applied directly to the data for two reasons. First, the differences among adjacent lists are not monotonic. Second, predictions based on different pairs of lists may differ.

### 4.1 Non-Monotonic Differences

The model shown in Fig. 3 is based on the fact that in both lists the number of citations is a monotonic function of the rank, and also assumes that the difference in citations between lists is a monotonic function of the rank. But this assumption is false. As researchers move up or down in the list the differences between counts at adjacent ranks change, leading to fluctuations of the differences between the lists. Fig. 4 shows that this is indeed the case in reality.

The lack of monotonicity is problematic because then the notion of a rank where a certain difference is "natural" is ill-defined. To cope with this situation, we smooth the data

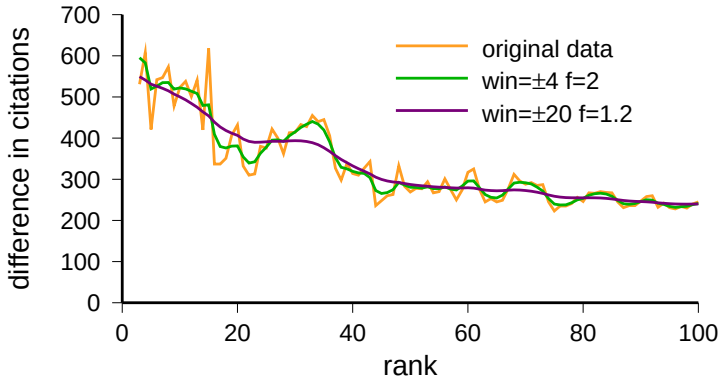


Figure 4: *Original differences of citation counts between the 11/2001 and 5/2002 lists, and the smoothed data.*

by applying a low-pass filter. This is achieved by averaging each value with its neighbors in a certain window, using exponentially decreasing weights for values that are farther away. While not guaranteed to produce a monotonic sequence, the results are nevertheless much better. With a large enough window, and small enough weight factor, the sequence can indeed approach monotonicity (see Fig. 4).

To finally settle the issue of a "natural" rank for a given difference, we conduct a bi-directional search of the smoothed list of differences. We note the highest and lowest points in the list at which the desired difference was observed, and use the midpoint between these two ranks as the asymptotic rank.

## 4.2 Non-Unit Time Intervals

The model of how the number of citations changes with time is based on $\Delta_p$ being the change for person $p$ during one time unit, and specifically from time $t-1$ to time $t$. We use one month as the time unit for our predictions. But the data comes from lists that are published at irregular intervals of several months. We therefore need to deduce the value of $\Delta_p$.

Recalling the definition of $\alpha_p = C_p(t)/C_p(t-1)$, we can write the identity

$$\Delta_p = C_p(t) - C_p(t-1) = C_p(t)\ (\alpha_p - 1)$$

This reduces the problem to that of finding $\alpha_p$, which is easily solvable, since by definition $C_p(t+\delta) = C_p(t)\alpha_p^\delta$, so

$$\alpha_p = \sqrt[\delta]{\frac{C_p(t+\delta)}{C_p(t)}}$$

## 4.3 Combining Lists

The theory developed above is based on charting the differences in citations between two ranking lists. At the time of writing, we have five usable ranking lists spanning the period

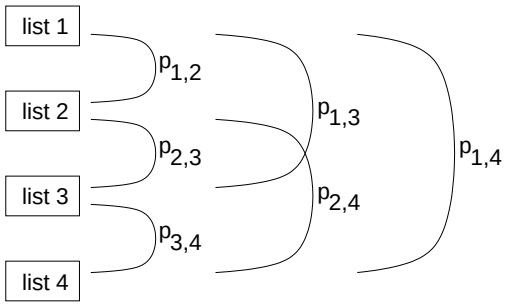


Figure 5: *Pairs of lists used to derive predictions.*

from April 2001 to May 2003. The methodology can therefore be applied to various pairs of lists. Regrettably, this leads to (sometimes grossly) different results.

The solution we adopted is to use a weighted average of the results obtained using different pairs of lists. This leads to two questions: which pairs to use, and how to do the weighting.

Consider a set of four lists. This leads to the pairs shown in Fig. 5. There are three pairs of distance 1, two of distance 2, and one of distance 3. Any prediction, e.g. the asymptotic rank of a certain person, or the predicted rank for a certain person in a certain month, can be calculated using each of these pairs. Denote the prediction obtained by lists $i$ and $j$ by $p_{i,j}$. This will be given a weight $w_{i,j}$ in the weighted average.

We checked five different weightling schemes, that take the distance and the recency into account. The base case gives exactly the same weight to all the predictions. If $k$ lists are available, there are $\binom{k}{2}$ ways to choose pairs of lists. The weight of each one is then

$$w_{i,j} = \frac{1}{\binom{k}{2}}$$

An alternative is to define a weighting factor $f$, and increase the weight by a factor of $f$ as the distance grows (we used a factor of 2 when testing this idea). This is based on the observation that adjacent lists sometimes display large fluctuations in ranking, which lead to wild predictions. The farther away the lists are, the more stable the results. The weight is then equal to

$$w_{i,j} = \frac{1}{k-(j-i)} \cdot \frac{f^{j-i-1}}{\sum_{\ell=0}^{k-2} f^{\ell}}$$

To give higher weight to more recent data, we assign the relative weight $i+j$ to $p_{i,j}$, based on the assumption that the lists are numbered in chronological order. This leads to

$$w_{i,j} = \frac{i+j}{\sum_{i'=1}^{k-1} \sum_{j'=i'+1}^{k} i'+j'}$$

These ideas were combined in various ways and compared. Evaluation was done by creating empirical confidence intervals as described below, based on 4 available lists, and counting how many of the results of the fifth available list fall into the confidence interval in each case. The results were that all weighting schemes gave very similar results. Fluctuations depended more on which lists were used than on the weighting scheme.

| *lists* | *hits* |
|---|---|
| 2, 3, 4 | 4201 (45.0%) |
| 2, 3, 5 | 5416 (58.0%) |
| 2, 4, 5 | 4715 (50.5%) |
| 3, 4, 5 | 3660 (38.8%) |
| 2, 3, 4, 5 | 6211 (66.6%) |

Table II: *Results of prediction accuracy using different sets of lists. Hits are actual ranks from list 6 that fall within the empirical confidence intervals predicted by the given lists. Percents are relative to the names that appear in all lists, which is around 9,330 in most cases.*

Recall that we have a total of 6 lists, of which list number 1 is not used, and list number 6 is used for testing the predictions. This leaves four candidate lists to make predictions. Table II shows the quality of results obtained using different subsets of these four lists. The results indicate that it is best to use all available data, i.e. all four lists. When using only three lists, it is more important to span a large difference (using sets $\{2, 3, 5\}$ or $\{2, 4, 5\}$) than to use the most up-to-date data (set $\{3, 4, 5\}$). But this conclusion should be reconsidered in the future when much more data spanning more time is available.

## 4.4 Empirical Confidence Intervals

The above theory does not include facilities for calculating confidence intervals. However, we can make an empirical assessment of our accuracy. We do so by defining a "confidence interval" about the predictions that empirically contains a large fraction of the true results.

There are two reasons why predictions may not be accurate. One is that the future sometimes contains surprises. The other is that the original data is not very good. To distinguish between these cases to some degree, we base the confidence interval on the dispersion of the data. In particular, we perform a simple regression analysis of the last four data points, leading to a linear predictor of rank as a function of time. We then find the differences between the actual data points and this line. The root-mean-square of these differences, denoted $d$, forms the basis for the confidence interval. If the points all fall near the line, the data for this researcher seems to be good, and the confidence interval will be small. If they are dispersed, the data is uncertain, and the confidence interval will be larger.

At this stage we have two values for each researcher: the predicted rank $p$ and the dispersion $d$. As in conventional calculation of confidence intervals, we multiply $d$ by a constant $f$ that ensures that a certain desired fraction of the results indeed falls within the confidence interval. Specifically, $f$ is selected as follows. Using 4 lists we can make predictions based on the first three, and check their quality using the actual data from the fourth list. We set $f$ to a value such that the confidence interval will include the true value in approximately half of the cases. Using lists 2, 3, and 4, and testing with list 5, led to selecting $f = 2$, which gave a hit rate of 56.5%.

As $d$ may be very small, we also added a constant factor that provides a minimal confidence interval independent of all other considerations. We also found it appropriate to make

part of this value proportional to the rank $r$, as mobility (and uncertainty) at the top of the list is much lower than at lower ranks. All these considerations taken together led to using the following formula for the initial confidence interval:

$$p \pm \left[ 2 + 25\frac{r}{10,000} + 2d \right]$$

Another question is how the confidence interval should evolve with time. This can be resolved by extending the same methodology as outlined above: making predictions based on only three lists, setting $f$ to include about half of the data points from the fourth list, and setting the growth parameter to include about half of the data points from the fifth list. Alternatively, we can tentatively set a growth rate proportional to $d$ for each researcher, and use a formula similar to that of the initial confidence interval but with smaller constants.

Examples of the results of this whole process are shown in Fig. 6. Note how the predicted ranking converges to the predicted natural rank, even if the initial rate of change is very high.

# 5 Evaluation

## 5.1 Quality of the Model

The problem with evaluating the quality of predictions is that we need to wait several years to see the true outcome for comparison. A sound evaluation is therefore not possible at this time.

What we can do is an initial evaluation for a short time range. We can create predictions based on all the available lists but the last one, and compare this with the actual values in this last list, which is several months later (in Section 4 we used a similar procedure to create the empirical confidence intervals). The results are shown in Figure 7. The relative error in ranking is typically less than 10%, and the absolute error is much smaller than 500 (in a list of 10,000).

As for the future, it should be realized that our predictions have inherent limitations. The predictions are based on current data, and assume that similar behavior will persist in the future. It is not possible to predict breakthroughs that will propel a certain researcher to much higher levels. And even without breakthroughs, citation counts may fluctuate greatly. Two striking examples of how predictions change when additional data is given are shown in Figure 8.

Moreover, the current model assumes a stationary rate of accumulating citations. This is probably not the case for truly exceptional researchers, that are destined to reach the top of the list. As both the number of citations and the rate of growth are very high at the top of the list, it is more reasonable to assume that the rate of accumulating citations will also grow with rank. In principle this can be modeled by the second derivative of the citation accumulation process, but current data seems to be too noisy for valid modeling. Nevertheless, the model does make it possible to identify researchers that will most probably rise to ranks that are much higher than their current rank.

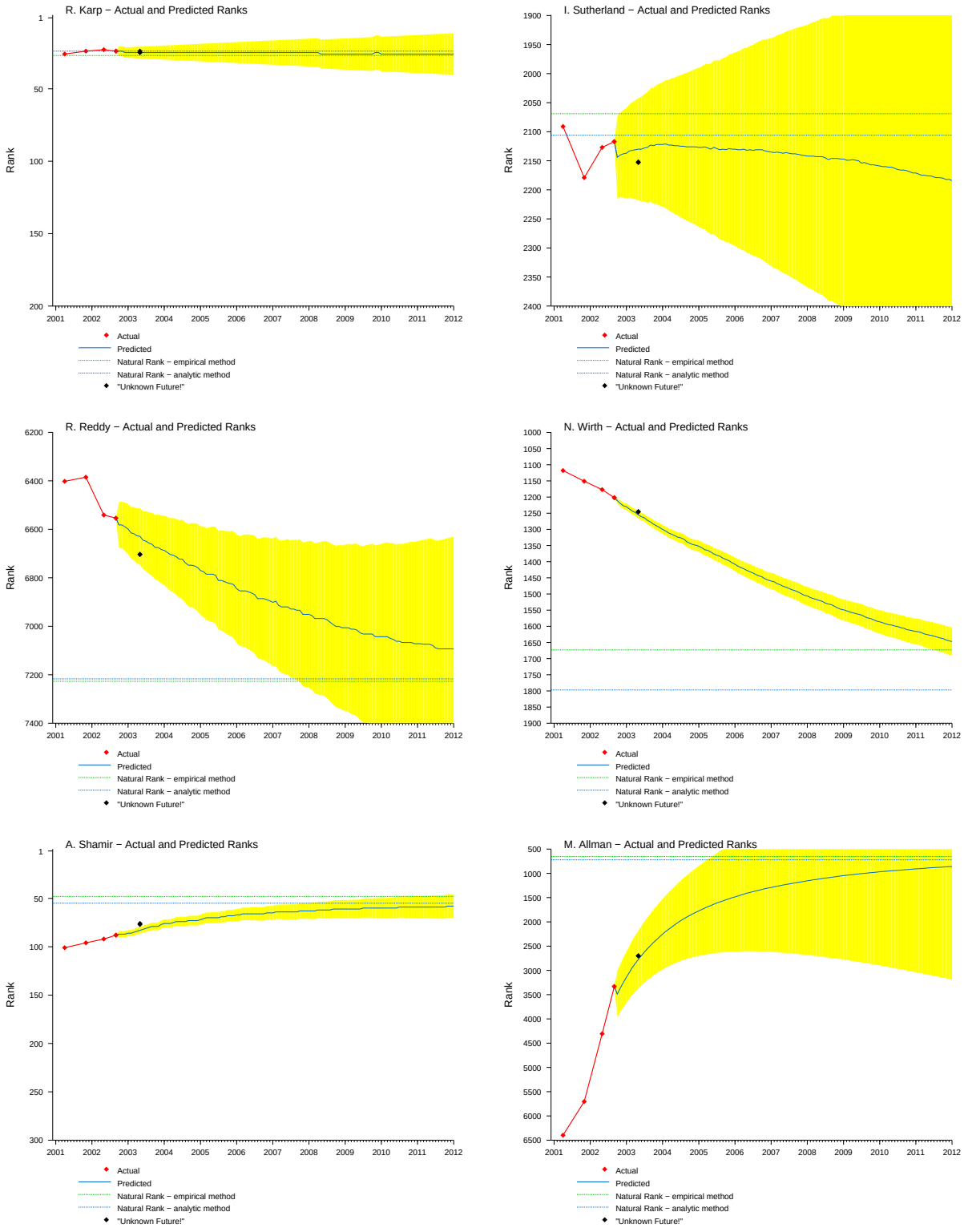


Figure 6: *Examples of predictions generated by our model based on four lists, compared to the actual data from a fifth list.*

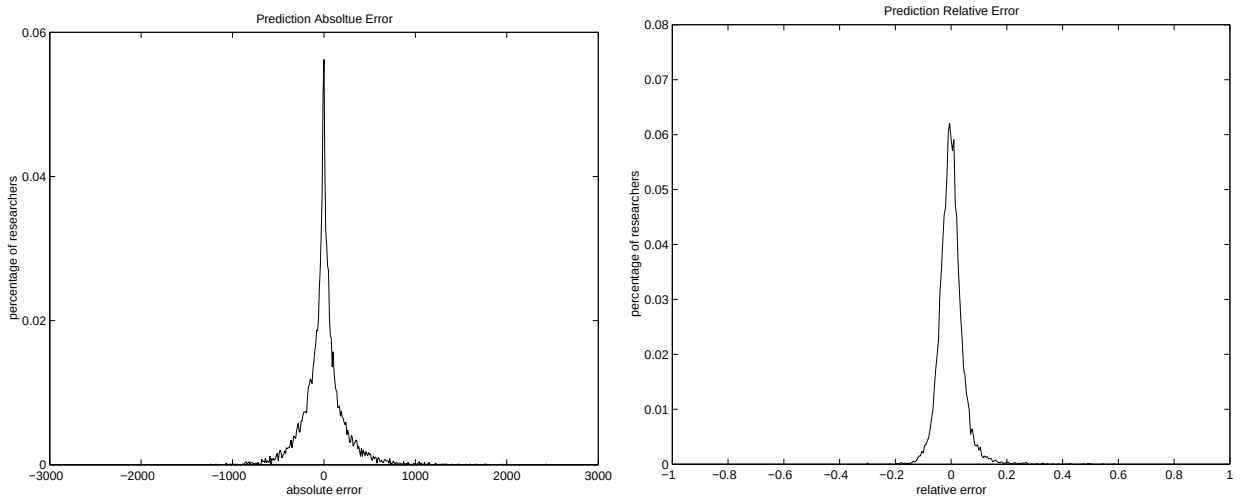


Figure 7: *Absolute and relative errors in estimation of rank based on four lists, as verified using a fifth list.*

## 5.2 Ranking by Citations

More generally, there are problems with the application of citation data to the ranking of scientists. First, citations are not necessarily a good metric. For example, Table III shows the ranking of ACM's Turing Award winners from its inception in 1966 to 2002. This is widely considered the most prestigious award in Computer Science. While some recipients indeed rank at the top of the list, indicating a good match between citations and standing, others are ranked quite low, and no less than 5 don't even make it into the list of the top 10,000 most cited authors. For example, Douglas Engelbart is inventor of the computer mouse and has contributed to other aspects of human-computer interaction. This has had great impact, but garnered few citations. Our predictions indicate that in upcoming years several others will also be dropped off the list.

The literature on citation indexing is rich with debate regarding problems with the interpretation of citations as a metric for quality and possible solutions. The MacRoberts provide an overview of the main objections, based on an analysis of why and when scientists cite (or do not cite) each other (MacRoberts and MacRoberts, 1989). This is corroborated by recent empirical data, which suggests that sources that are considered very bad are cited more than those considered only moderately bad, maybe as an example of what should not be done (Nicolaisen, 2002).

At the other end of the spectrum, it is plausible that breakthrough papers do not receive the citations they deserve due to the "citation food chain". A breakthrough article may initially be cited directly, but later it may be replaced by a survey of the whole field as the favorite citation. Take the theory of NP-completeness as an example. The original FOCS 1971 paper contributed 444 of Cook's 1926 total citations, earning him the respectable rank of 334 in the May 2003 list. Karp's followup paper from the next year got 571 of Karp's 4951 citations, which together propelled him to rank 24. But the most common reference is the book by Garey and Johnson, which, at 3392 citations, is the most cited source in the

| *year* | *recipient* | *cites* | *rank* |
|---|---|---|---|
| 1966 | A.J. Perlis | 211 | 8811 |
| 1967 | Maurice V. Wilkes | 50 | **n/a** |
| 1968 | Richard Hamming | 237 | 7744 |
| 1969 | Marvin Minsky | 1205 | 849 |
| 1970 | J.H. Wilkinson | 1119 | 955 |
| 1971 | John McCarthy | 3085 | 108 |
| 1972 | E.W. Dijkstra | 2915 | 130 |
| 1973 | Charles W. Bachman | 23 | **n/a** |
| 1974 | Donald E. Knuth | 5793 | 12 |
| 1975 | Allen Newell | 2450 | 202 |
| 1975 | Herbert A. Simon | 3962 | 49 |
| 1976 | Michael O. Rabin | 1718 | 433 |
| 1976 | Dana S. Scott | 2440 | 206 |
| 1977 | John Backus | 358 | 4768 |
| 1978 | Robert W. Floyd | 802 | 1611 |
| 1979 | Kenneth E. Iverson | 73 | **n/a** |
| 1980 | C. Antony R. Hoare | 4758 | 29 |
| 1981 | Edgar F. Codd | 950 | 1233 |
| 1982 | Stephen A. Cook | 1926 | 334 |
| 1983 | Ken Thompson | 1146 | 915 |
| 1983 | Dennis M. Ritchie | 396 | 4216 |
| 1984 | Niklaus Wirth | 946 | 1245 |
| 1985 | Richard M. Karp | 4951 | 24 |
| 1986 | John Hopcroft | 4542 | 34 |
| 1986 | Robert Tarjan | 6525 | 7 |
| 1987 | John Cocke | 1074 | 1017 |
| 1988 | Ivan Sutherland | 663 | 2152 |
| 1989 | William (Velvel) Kahan | 413 | 3973 |
| 1990 | Fernando J. Corbato' | 34 | **n/a** |
| 1991 | Robin Milner | 7900 | 4 |
| 1992 | Butler W. Lampson | 1643 | 471 |
| 1993 | Juris Hartmanis | 742 | 1817 |
| 1993 | Richard E. Stearns | 380 | 4434 |
| 1994 | Edward Feigenbaum | 363 | 4684 |
| 1994 | Raj Reddy | 270 | 6703 |
| 1995 | Manuel Blum | 1704 | 442 |
| 1996 | Amir Pnueli | 5212 | 19 |
| 1997 | Douglas Engelbart | 113 | **n/a** |
| 1998 | James Gray | 3945 | 50 |
| 1999 | Frederick P. Brooks, Jr. | 908 | 1332 |
| 2000 | Andrew Chi-Chih Yao | 2019 | 304 |
| 2001 | Ole-Johan Dahl | 505 | 3094 |
| 2001 | Kristen Nygaard | 498 | 3161 |
| 2002 | Ronald L. Rivest | 6930 | 5 |
| 2002 | Adi Shamir | 3492 | 76 |
| 2002 | Leonard M. Adleman | 1746 | 418 |

Table III: *Ranking of Turing Award winners according to the May 2003 CiteSeer list. Unranked winners have less citations than the 185 needed for a rank of 10,000.*

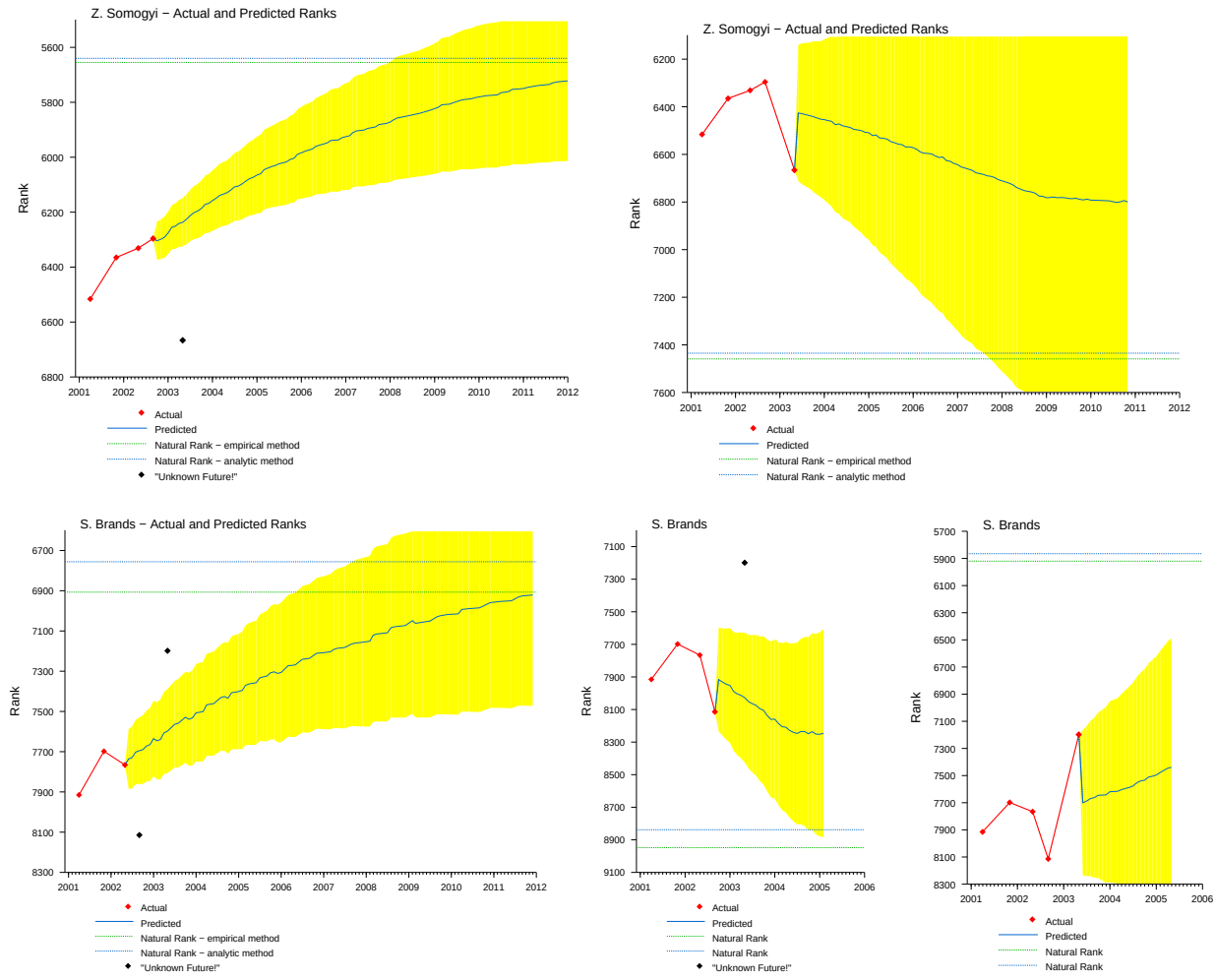


Figure 8: *Examples of extreme cases of how predictions change when additional data becomes available, specifically when the data from a new list does not agree with that of the previous lists.*

CiteSeer database[1]. It also contributes a sizable fraction of Johnson's 12119 citations (1st rank) and Garey's 6732 citations (6th rank).

It may also happen that a breakthrough is so successful, that it ceases to be cited because it has become common knowledge — a phenomenon called "obliteration" by Garfield (Garfield, 1979). As an example, consider Codd's 1970 paper in which he introduced relational databases. This paper now has some 410 citations. This does not begin to reflect the fact that practically all work on databases today is based on the relational model. Ullman's textbooks on database systems have 1364 citations — also a small number relative to all work being done on databases. In citation databases in general the top ranked items are textbooks (Goodrum et al., 2001), and the top ranked researchers in CiteSeer are mainly

[1]It should probably be even higher: the "researcher" C. Intractability, which has 872 citations, is most probably a mis-parsing of *Computers and Intractability*, the title of Garey and Johnson's book.

authors of textbooks.

Another problem with ranking by citations is the question of coverage and quality of the data. The number of citations of top papers is bounded by the size of the field: a field in which fewer papers are published necessarily produces fewer citations. Thus ranking papers from different fields is problematic. Whitley has shown that well-established citation indexing services do not completely overlap, indicating that each one misses a sizeable fraction of the literature (Whitley, 2002). These can be called errors of omission. Another type of obvious errors results from authors that have the same initials or even the same names, and cannot be separated.

The CiteSeer data, being collected by automatic means, also suffers from other types of errors. One type is parsing errors, in which authors or works are not recognized correctly (this seems to be shared at least to some degree by the Science Citation Index). For example, rank 6814 in the list is occupied by S. Engineering, with 266 citations, rank 5340 is occupied by C. University, with 328 citations, A. Computer is ranked at 4913 with 350 citations, C. Systems achieved the respectable rank of 2781 with 547 citations (9 other members of the Systems family, with different initials, are also represented), S. Http is at rank 2548 with 583 (and there are 12 others of the Http family too), and C. Intractability is at rank 1403 with no less than 872 citations. A possible sign that the situation is improving is the fact that the rank of A. Introduction has been steadily declining, and he did not make it into the top 10,000 list of September 2002. Another problem is repetition errors, in which marginally different pre-publication versions of the same document are found, leading to double counting of references. Finally, being based on papers posted in personal home pages, CiteSeer may be susceptible to cultural differences among researchers. For example, if systems-oriented researchers in general tend to their web pages more than theoreticians, they will probably be found to have more citations.

# 6 Conclusions

Using citation data for quality ranking is risky. Predictive ranking is even riskier. At best, our model can predict how many citations authors will accrue relative to others; it does not support or provide any interpretation of why one author gets more or less citations. The goal of this work is not to promote the use of citations as a means for ranking (and subsequent hiring and promotion decisions). Citation data may at best be one of several inputs to important decisions, and should definitely not be the decisive one. An important part of our work aims to draw attention to the problems, complexities, and inconsistencies encountered when using such data.

As for our proposed model for predictive ranking, it seems to be a viable first step. Based on observations of the CiteSeer data and on the model, we can make the following remarks:

- There is a good correlation between citations and impact or recognition, but citations cannot be used as the decisive and conclusive metric for performance. Some highly recognized researchers have very few citations. Most researchers with very many citations got them by writing widely used textbooks.

- The model enables the identification of rising stars, and a rough estimate of how high they will rise in the list. This is based on fast and consistent improvement of the rank. However, the accuracy of long range predictions is questionable, especially since the dynamics near the top of the list can be expected to be different from those in the lower half.
- Over most of the list a small change in citations can lead to a sizeable change in ranking. This makes the predictions vulnerable to noise in the data. Predictions based on data that shows inconsistent behavior in the past are probably not reliable.

The model definitely has its limitations. The two chief ones are the assumptions that citations will continue to accrue at essentially the same rate, and that all ranks accrue citations at a common rate. Many interesting extensions and additions are possible. In particular, it would be fruitful to check the compatibility of this descriptive model with a detailed mechanical model of the whole citation process, including the population of researchers, the progress of their careers (Huber, 2002), the differences between the productivity of top researchers and "normal" researchers, and the accumulation of citations to individual papers (Burrell, 2002). A first step could be comparing with citation predictions provided by CiteSeer, or using the somewhat more detailed model of (Geller et al., 1981). Another important extension would be to model the whole list at once, rather than modeling individual researchers against a backdrop of constant growth at all ranks. This will verify the appropriateness of the multiplicative model.

An interesting observation is that the normal situation is for the rank to worsen with time. This is due to the growth of the population of researchers, and to the advance of new high-ranking ones: each new star that moves up by $x$ steps in the ranking causes $x$ others to move down by one step. Modeling this may enable one to distinguish between researchers whose rank is degraded "normally" (signifying stability rather than deterioration) from those whose rank is degraded faster, e.g. due to a shift in their careers. Moreover, it suggests that ranking dynamics should show that the rank first improves until it reaches a peak, and then declines. This implies a completely different model than ours, in which ranks converge asymptotically.

# References


Burrell, Q. L. (2002), "The $n$th-citation distribution and obsolescence". *Scientometrics*, 53(3):309–323.

"NEC Research Institute CiteSeer". URL http://citeseer.nj.nec.com/.

Garfield, E. (1979), *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons.

Geller, N. L., de Cani, J. S., and Davis, R. E. (1981), "Lifetime-citation rates: A mathematical model to compare scientists' work". *J. Am. Soc. Inf. Sci.*, 32(1):3–15.

Goodrum, A. A., McCain, K. W., Lawrence, S., and Giles, C. L. (2001), "Scholarly publishing in the Internet age: A citation analysis of computer science literature". *Information Processing & Management*, 37(5):661–675.

Holmes, A. and Oppenheim, C. (2001), "Use of citation analysis to predict the outcome of the 2001 research assessment exercise for unit of assessment (UoA) 61: Library and information management". *Information Research*, 6(2). URL InformationR.net/ir/6-2/paper103.html.

Huber, J. C. (2002), "A new model that generates Lotka's law". *J. Am. Soc. Inf. Sci.*, 53(3):209–219.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999), "Digital libraries and autonomous citation indexing". *Computer*, 32(6):67–71.

MacRoberts, M. H. and MacRoberts, B. R. (1989), "Problems of citation analysis: A critical review". *J. Am. Soc. Inf. Sci.*, 40(5):342–349.

May, R. M. (1997), "The scientific wealth of nations". *Science*, 275:793–796.

Nicolaisen, J. (2002), "The J-shaped distribution of citedness". *J. Doc.*, 58(4):383–395.

"ScienceWatch". URL http://www.sciencewatch.com/.

Whitley, K. M. (2002), "Analysis of SciFinder Scholar and Web of Science citation searches". *J. Am. Soc. Inf. Sci.*, 53(14):1210–1215.