The growth and memory of science

Kevin W. Boyack

Sandia National Laboratories* , P.O. Box 5800, Albuquerque, NM 87185

E-mail: kboyack@sandia.gov

Alex Bäcker

Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185

and

California Institute of Technology, MC 139-74, Pasadena, CA 91125

E-mail: alex@caltech.edu

* **Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.**

Abstract

It is a cliché that the pace of scientific advance is accelerating. All other things equal, this belief would lead to the prediction that, on average, scientific papers build on the findings of ever younger papers. In addition, changes in the distribution of scientific literature may have shifted the network of papers used by the average scientist towards those papers available electronically. We examined the net effect of these forces on a macro scale by looking at the evolution of the age distribution of references over time. We found that mean reference ages have grown steadily over the last 20 years, and that this growth can be explained by changes in the growth dynamics of science. We show that the presumed exponential growth of scientific output ended before 1960. This growth suffered an abrupt increase in slope *ca.* 1960, and has followed roughly linear growth since. Both this change and linear growth track growth in academic R&D investment closely. After growth is accounted for, our findings reveal a decrease in the memory of science starting in 1982 that has reversed in the last 10 years. Finally, we show that the aging of 80+-year-old literature seems to have ceased, implying that classics endure indefinitely in the collective memory of science.

The last decade has seen the popularization of the Internet, greater availability of scientific literature in electronic form, and novel ways to store and catalog information. Our collective perception is that the pace of scientific advance is accelerating. Have these transformations produced any changes in scientific memory, as measured by referencing of prior literature? We hypothesized that the increasing speed of scientific developments and enhanced access to the newest publications might have decreased the probability of citations to older publications. In contrast, and to our surprise, we found the opposite.

Citation networks provide a powerful tool to probe the history of scientific disciplines. The set of references of all papers published in any given year constitutes a mirror of the body of science that is in the collective conscience of scientists at the time. Citation age distributions can thus be seen as a window on the memory of science, and can help us quantify the overall effects of transformations such as those listed above. Study of the age distribution of references has a long history, starting with the influential analysis of Price[1] and continuing with the work of Griffith[2,3] and van Raan[4]. Using data from the 1961 Science Citation Index (SCI), Price noted that citations increased exponentially with cited year. Using the assumption that citation probabilities were independent of age for papers older than 15 years, the citation curve thus became a proxy for the publication curve. Although the actual numbers of papers published in cited years was not measured, this study by Price[1] validated the widely-held notion that science grows exponentially, a belief that persists today[5-11]. Applying Price's method, using reference counts as a proxy for paper counts, to current reference data from the Science Citation Index Expanded (SCIE), Figure 1 shows that growth in scientific output was indeed exponential in the aggregate from 1720 up to at least 1900[12].

In order to test our hypotheses about changes in the memory of science, we first calculated mean reference ages by citing year using data from a 25-year segment of the SCIE. Although other publication data sources abound, the SCIE is the most suitable data source for this type of study in that 1) growth in the SCIE is commonly used as a proxy for growth in science [6], 2) it is the only extensive source of citation information covering the most important publications across all major scientific disciplines, and 3) roughly 75% of the references are to sources within the SCIE, indicating that it does cover the majority of the best peer-reviewed science. Using the over 300 million references from 15 million articles indexed in the SCIE from 1977 to 2001, we found that mean reference age increased steadily over that time period for references in all age brackets examined (Figure 2a). Consistent with this increase, the percentage of references to papers 5 years old or less has been decreasing, from 47.4% of references in publication year 1980 to 43.0% of references in publication year 2001. A recent finding that citing half-life has increased from 9.3 years in 1975 to 10.5 years in 2002 is consistent with our findings[13]. The distance between the curves changes most between the 15 year and 50 year points, indicating that much of the overall change in mean reference age stems from citations to papers 16-50 years old.

Mean reference age is affected by both growth[14] and aging[15]. Growth in the numbers of papers published each year affects the numbers of papers of each age available to be cited. Aging can be described by the curve relating how likely it is for a paper to be cited as a function of its age. Growth and aging combine to determine how many papers of any given age are expected to be cited in any given year. To assay whether the increase in mean reference age could be explained by aging and growth, we calculated expected mean reference ages taking growth into

account and using realistic retrospective citation rates (i.e. aging)[16]. This required measuring the historic growth of the collective body of scientific works using actual numbers of scientific papers published (articles, letters, notes, and reviews; hereafter ALNR) rather than , as was used by Price, numbers of references to past publications.

The number of papers indexed in the SCIE for 1945 to present[17] (Figure 3) reveals a sharp increase in slope *ca.* 1960. We verified the generality of this effect by observing that similar changes *ca.* 1960 can be seen in publication data from *Chemical Abstracts*, *Physics Abstracts*, and *Electrical and Electronics Abstracts*[18]. One might assume that this change was due to changes in indexing practices (i.e., use of mainframe computers) and thus would be common to the different indexing services. However, while *Chemical Abstracts* has been indexing literature consistently since the early 1900s, indexing for the SCI did not start until the early 1960s, and earlier data (1945-1960) were only indexed and added recently. Thus, changes in indexing practices cannot account for the change ca. 1960. Thus, the publication pattern of Figure 3 is likely representative of actual publication numbers rather than just of the specific indexing choices of one indexing service. These data are intriguing in that they suggest that growth in science has been linear over the past 40 years, rather than exponential as is widely assumed. Note that although a linear increase is slower than an exponential one in the limit of infinite time, in this case the change to linearity was accompanied by a large boost in the slope of scientific growth.

We next looked for the causes underlying such a dramatic shift in the trend of global scientific output. Figure 3 relates scientific output, measured in number of papers, to R&D

investments[19] by the U.S. and by all of the *Organization for Economic Co-operation and Development* (OECD) countries. Despite cynicism in the literature about the capability of funding to change science ([20, p. 80] and Price, as reported by Griffith [2]), Fig. 3 shows that increases in R&D investment track the increase in the number of papers over the same time period. R&D performed by U.S. academic institutions (measured in constant dollars) increased by a factor of 8.9 from 1960 to 2000, while the number of papers increased by precisely the same factor, 8.9, over the same period. Academic R&D provides the best measure of funding for publication-driven research given that a majority of papers (~60%) are published by academia[21]. The roughly 70% of U.S. R&D spent by industry accounts for less than 3% of published papers. Investment in the U.S. alone is unlikely to account for the global increase in science production. The OECD (G7 plus Russia) investment curve, however, parallels that of the U.S. in the 1980s and 1990s, suggesting their increases may track each other. We considered insertion of a 2 year time-lag to account for the time between funding and publication (by moving the funding curves backward in time), but this did not change the results significantly. Thus, it appears highly likely that R&D investment has driven growth of science since 1960.

A recent study of the growth dynamics of university research has shown that the distributions of funding and papers follow the identical functional forms[22]. By showing that research output follows funding changes over time, our results provide a potential mechanism to explain how such identical distributions came to be. The fact that the curve of scientific output shows a more steady increase than the R&D investment curves, which show more pronounced idiosyncratic historical effects, such as the economic difficulties of the early 1970s and 1990s, is consistent with a long-term (and thus low-pass filtered) effect of R&D investment.

It is remarkable that the aggregate growth in scientific output has been so smooth and predictable since the early 1960s. This is the case despite perceived changes in publication practices over time (e.g. "paper inflation", a name coined by some who believe that papers in times past were larger or more substantial than today's), and different rates of evolution in individual technical domains, both increasing[23] and decreasing[24], over the last half century. This may reflect a fundamental phenomenon that scientists may switch disciplines, but, for the most part, remain scientists for a lifetime. Thus, if scientific output is constrained by the number of scientists[20], changes in the output of individual scientific disciplines are bound to balance each other out to match the growth of the scientific workforce.

Armed with these data on the growth of science, we were able to ask whether the increase in mean reference ages was due to the changes in growth dynamics. We calculated the mean reference age curves that would have been expected accounting for both growth and aging[25]. Expected mean reference ages, accounting for growth and aging, increase with time above and beyond the increase seen in the actual mean reference curves (Fig. 2a), suggesting that growth and aging account for the observed increases in mean reference age, and that the *intrinsic, or growth-normalized,* memory of science has been decreasing.

To examine this change in the aging curve more closely, we defined and calculated growth-normalized mean reference age – GNMRA[26]. GNMRA has several desirable properties. It is independent of the number of papers published in any given year. It exhibits an increase if mean reference ages go up compared to what would be expected from growth. Finally, GNMRA

coincides with mean reference age if there is no change in growth-normalized aging curves. Because of this, GNMRA can be used to compare citation practices of different citing years. GNMRA decayed steadily between 1982 and 1995, and has been increasing since (Fig. 2b), indicating a shift toward citing more recent publications during the 1980s and early 1990s followed by a reversal of the trend since 1995.

GNMRAs condense the entire aging curve to a single number. To examine the change in citation practices in more detail, as a function of reference ages, we set out to compute normalized citation counts for all pairs of citing and cited publication years as

$$C(T,t) = R(T,t) / [R(T)\, N(t)] \ ,$$

where $R(T,t)$ is the number of references from publication year T to cited year t, $R(T)$ is the number of references from year T to all previous years, and $N(t)$ is the number of papers published in any year t. This measure accounts for the growth in references per paper as well as for the growth of publication numbers[27]. These normalized aging curves reveal that the aging of 80+-year-old literature seems to have ceased (Fig. 4), marking a shift from the so-called obsolescence (or aging) of older literature – the finding that documents older than about 50 years are cited less often as the years go by—found by Griffith[2,3] using data from 1975 and 1986. They also show that aging curves corresponding to different publication years diverge at ages corresponding to the 1960 inflection in the growth curve, underscoring the importance of this event in shaping modern citation distributions.

In sum, we found that, surprisingly, the growth of science is no longer exponential, as was previously and widely assumed, but has been roughly linear since at least 1960. The cessation of exponential growth in scientific output mirrors a similar end, also in the 1960s, in the exponential growth of the number of scientific periodicals[28]. Second, the rate of growth of scientific output rose sharply *ca.* 1960. Third, this linear growth in scientific output closely tracks the increase in R&D funding. Interestingly, this singular change in the pattern of scientific growth closely follows the launch of Sputnik in 1957, an event that put science in the eye of policymakers. The acceleration of scientific output *ca.* 1960 is in marked contrast with the reduction in the number of scientific periodicals since the 1960s[28], a disparity that may indicate that the investment in science that has been fueling scientific growth in the last four and a half decades has gone mostly to enlarge research groups, not to diversify them. These findings raise intriguing questions: Is the cost of scientific advancement essentially constant in terms of the amount of advance per unit funding? What are the relative effects of labor and capital on scientific advance? Answers to these questions may help us to more effectively foster growth in science.

It is interesting to speculate on the causes underlying a shift in the dynamics of scientific growth from exponential to linear. Exponential growth is consistent with the notion of a each scientist training a constant mean number of apprentices who go on to train more scientists themselves, a model coherent with the academic beginnings of modern science. In contrast, saturation of the academic job market roughly coincident with the explosion of academic funding in the early 1960s would have led to increasing number of scientists leaving academia, thus ceasing to contribute to the growth of the scientist pool itself. Under these new dynamics,

scientific growth becomes a function of the size of the job market for scientists, which is itself a function of R&D funding, a scenario that is consistent with our findings.

We also found that despite the common notion that the pace of science and innovation is accelerating, not only is the pace of discovery growing only linearly, but scientists are also looking farther and farther back into the past, as measured by an increasing mean age of references in the scientific literature. This is due to the fact that science has ceased its exponential growth, and thus the fraction of total science constituted by the most recent year's publications is diminishing. After accounting for scientific growth, normalized citation ages were decreasing until recently, during the early phase of popularization of the WWW, but have been growing again in the last 10 years, a trend consistent with a recent surge in "deep" archives that provide online access to older literature. Aging of literature older than 80 years old seems to have ceased. Interestingly, this phenomenon allows for a discrete characterization of papers into those that cease to be cited within 80 years of publication, and classics, which achieve immortality. As the growth of science continues its course unabated, it is somewhat reassuring to see that classics persevere in the collective memory of science.

References

1. Price, D. J. D. Networks of scientific papers. *Science* **149**, 510-515 (1965).
2. Griffith, B. C. Derek Price's puzzles: Numerical metaphors for the operation of science. *Science, Technology & Human Values* **13**, 351-360 (1988).
3. Griffith, B. C., Servi, P., Anker, A. & Drott, M. C. The aging of scientific literature: A citation analysis. *Journal of Documentation* **35**, 179-196 (1979).
4. van Raan, A. On growth, ageing, and fractal differentiation of science. *Scientometrics* **47**, 347-362 (2000).
5. Price, D. J. D. *Little Science, Big Science* (Columbia University Press, New York, 1963).
6. Tabah, A. N. Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology* **34**, 249-286 (1999).
7. Avery, J. *Science and Society* (H. C. Orsted Institute, University of Copenhagen, 1995).
8. Chen, C. *Mapping scientific frontiers: The quest for knowledge visualization* (Springer-Verlag, London, 2003).
9. Friedman, J. I. University at Albany Graduate Commencement Ceremony Speech. *http://www.albany.edu/feature2001/5-18/friedman_commencement_speech.html* (2001).
10. Kurzweil, R. The law of accelerating returns. *http://www.kurzweilai.net/meme/frame.html?main=/articles/art0134.html* (2001).
11. David Goodstein (NCAR 48 Symposium 1994) has noted that the exponential growth in the number of scientists, as well as that in number of journals, ended ca. 1970. Yet the growth in the number of scientific articles is a different story: despite the fact that this

growth ceased being exponential well before 1970, its present value significantly exceeds the value it would have had had the previous exponential growth continued.

12. All studies of citation age distribution, including our results in Figure 1, rely on the assumption that the number of referenced papers as a fraction of the number of published papers does not change with cited year for old literature.

13. Marx, W. & Cardona, M. Blasts from the past. *Physics World* **17**, 14-15 (2004).

14. Weiss, P. Knowledge: a growth process. *Science* **131**, 1716-1719 (1960).

15. Goffard, S. J. & Windle, C. D. Life of scientific publications. *Science* **132** (1960).

16. $MRA(T) = \frac{\sum_{t=0}^{n} (T-t) * R(T,t)}{\sum_{t=0}^{n} R(T,t)}$, where R(T,t) is the number of references from publication year T, to cited year t. EMRA(T) = MRA(T) using the substitution R(T,t)=P(T-t)*N(T)*N(t) with P(T-t) an average from 1977-2001 P(T,t) data.

17. Our calculations required numbers of papers by year back into the 1800s. Such data are not readily available. Thus, we estimated publication numbers for the years 1830-1939 using an exponential, and using a linear interpolation from 1939-1945 to estimate the decrease in WWII scientific publication (see Fig. 1). A similar decrease in scientific production is known to have existed for the WWI years (1), but is ignored here because the numbers are very small compared to current publication rates. Actual numbers of ANLR were used for the years 1945-2001.

18. Gupta, B. M., Sharma, P. & Karisiddappa, C. R. Growth of research literature in scientific specialties: A modeling perspective. *Scientometrics* **40**, 507-528 (1997).

19. National Science Board. *Science and Engineering Indicators - 2002, Appendix tables 4-4 and 4-40* (NSF, Arlington, VA, 2002).

20. Menard, H. W. *Science: Growth and Change* (Harvard University Press, Cambridge, MA, 1971).

21. Leydesdorff, L. The mutual information of university-industry-government relations: An indicator of the triple helix dynamics. *Scientometrics* **58**, 445-467 (2003).

22. Plerou, V., Amaral, L. A. N., Gopikrishnan, P., Meyer, M. & Stanley, H. E. Similarities between the growth dynamics of university research and of competitive economic activities. *Nature* **400**, 433-437 (1999).

23. Hullman, A. & Meyer, M. Publications and patents in nanotechnology. *Scientometrics* **58**, 507-527 (2003).

24. Gravili, C., Pagliara, R., Vervoort, W., Bouillon, J. & Boero, F. Trends in hydroidomedusan research from 1911 to 1997. *Scienta Marina* **64**, 23-29 (2000).

25. To account for aging, we used the aging curves given by actual mean retrospective citation rates averaged over the entire time period of available data (1977-2001). In other words, we defined expected citation probabilities to be a function of the age of the referenced paper, but not of the citing year. Price (1) used a simple approximation to aging, suggesting that 70% of citations are randomly distributed among all existing papers, while the other 30% are selective to recent literature. Of this 30%, Price suggested half cite papers less than 6 years old.

26. $$GNMRA(T) = \frac{\sum_{t=0}^{n} (T-t) * R(T,t) * \dfrac{\dfrac{R(T,t)}{N(T) * N(T-t)}}{\dfrac{R(Tref, Tref-age)}{N(Tref) * N(Tref-age)}}}{\sum_{t=0}^{n} R(T,t) * \dfrac{\dfrac{R(T,t)}{N(T) * N(T-t)}}{\dfrac{R(Tref, Tref-age)}{N(Tref) * N(Tref-age)}}}$$

27. Van Raan's P(T,t) is similar to our C(T,t), but uses N(T) rather than R(T) in the denominator, and thus does not account for growth in number of references per paper through the years. The number of references per ALNR for the entire SCI has nearly doubled from 14.9 in 1977 to 27.8 in 2001.

28. Pendlebury, D. Science's go-go growth: Has it started to slow? *The Scientist* **3**, 14 (1989).

Acknowledgements

We thank G. Davidson, D. deB. Beaver, H. Gray and Caltech's Sloan-Swartz Center for Theoretical Neurobiology for discussions on the implications of our findings, and M. Changizi, K. Börner, C. Chen, W. Marx, P. Perona and S. Morris for comments on the manuscript. This work was supported by the Sandia National Laboratories Laboratory-Directed Research and Development Program, the DOE Office of Science's MICS Program and Caltech's Beckman Institute. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy.

Figure Captions

Figure 1. Age distribution of references from papers published in 2001 and included in ISI's Science Citation Index Expanded (SCIE). Scientific output grew exponentially (solid line) starting in the 1720s and until at least 1900. References after 1920 do not reflect number of publications, since they are subject to recency, or differential citation probabilities due to aging.

Figure 2. Mean reference ages (MRA) have been growing steadily (a, inset; 21 of 25 years analyzed showed higher MRA than the year before, $p < 0.0009$, two-sided binomial test), but have failed to keep up with the growth expected due to growth in publication numbers. a) Mean reference age using references of difference age brackets for different publication years. Publication years: 1977 (circles), 1989 (diamonds), and 2001 (squares). Actual values (open symbols) were calculated for different reference ages using all references less than *n* years old. Expected mean reference ages (filled symbols) were calculated using the formula in note [16] and cover a 150 year reference age window. b) Growth-normalized mean reference ages (GNMRA, see note [26]). The error bars show GNMRAs using references of only odd or only even ages.

Figure 3. Production of scientific papers rose sharply following the launch of Sputnik and correlates closely with linear growth in academic R&D investment, showing a cessation of previous exponential growth. Assuming exponential growth from 1980-2001 severely overpredicts the actual number of papers published prior to 1980. An assumption of exponential growth based on the 1950-1960 time period would underpredict the actual number of papers

since 1960. Correlation coefficients for linear and exponential fits to the number of papers from 1960-2001 are 0.995 and 0.903, respectively.

Figure 4. Classics last forever. Growth-normalized citation counts, *C(T,t)*, for different publication years as a function of reference age show that citation probabilities are constant after age 80. For the 80+ year-old literature, decreases in *C(T,t)* are seen from 1980 to 1987 and 1987 to 1994. However, the 2001 curve is at the same level as the 1994 curve, indicating that aging of century old literature has ceased. The filled circles indicate cited year 1960 on each of the four reference curves. Each curve separates from its predecessor above this point, consistent with the disruptive change in publication counts ca. 1960.

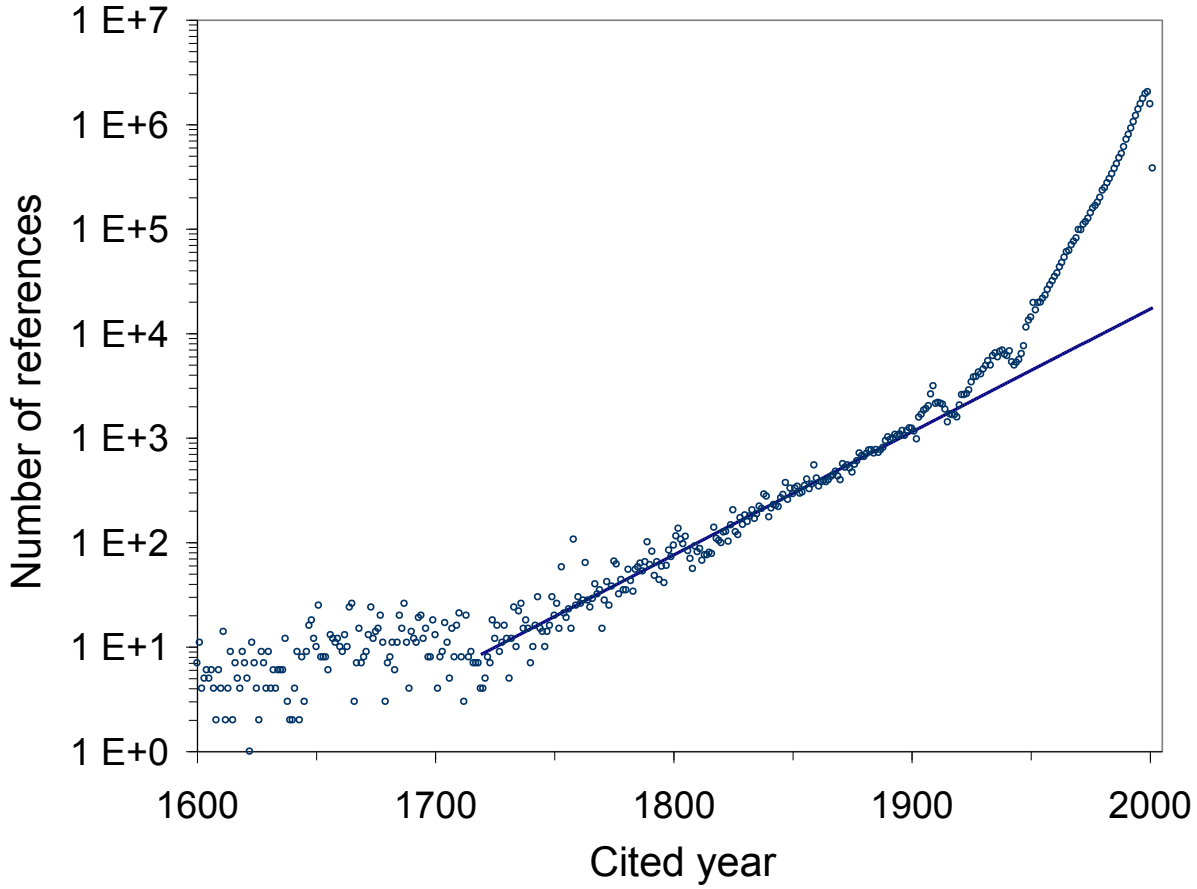
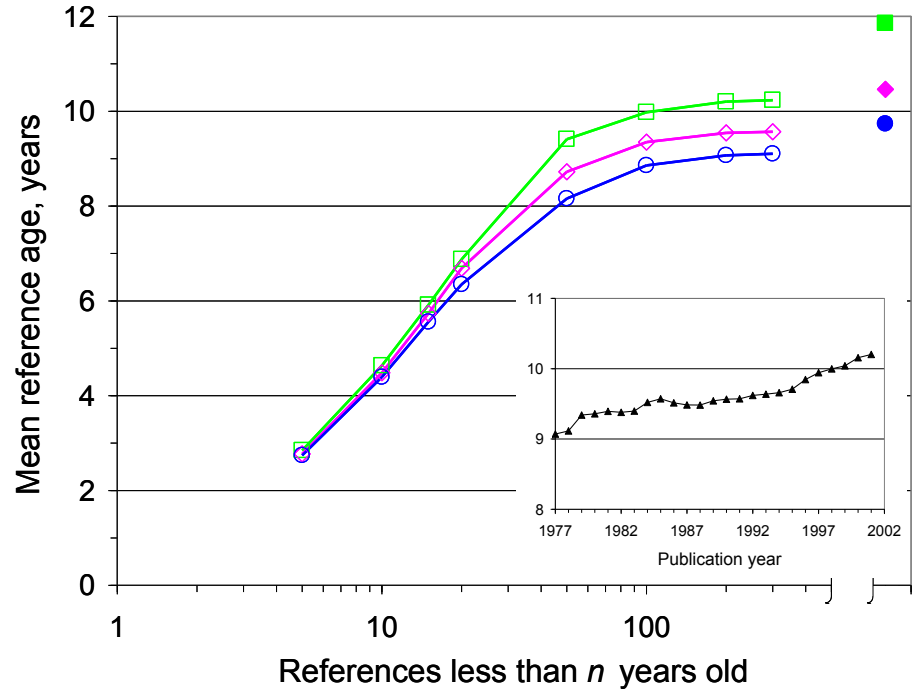

1 E+7
1 E+6
1 E+5
1 E+4
1 E+3
1 E+2
1 E+1
1 E+0
Number of references
1600
1700
1800
1900
2000
Cited year

Figure 1



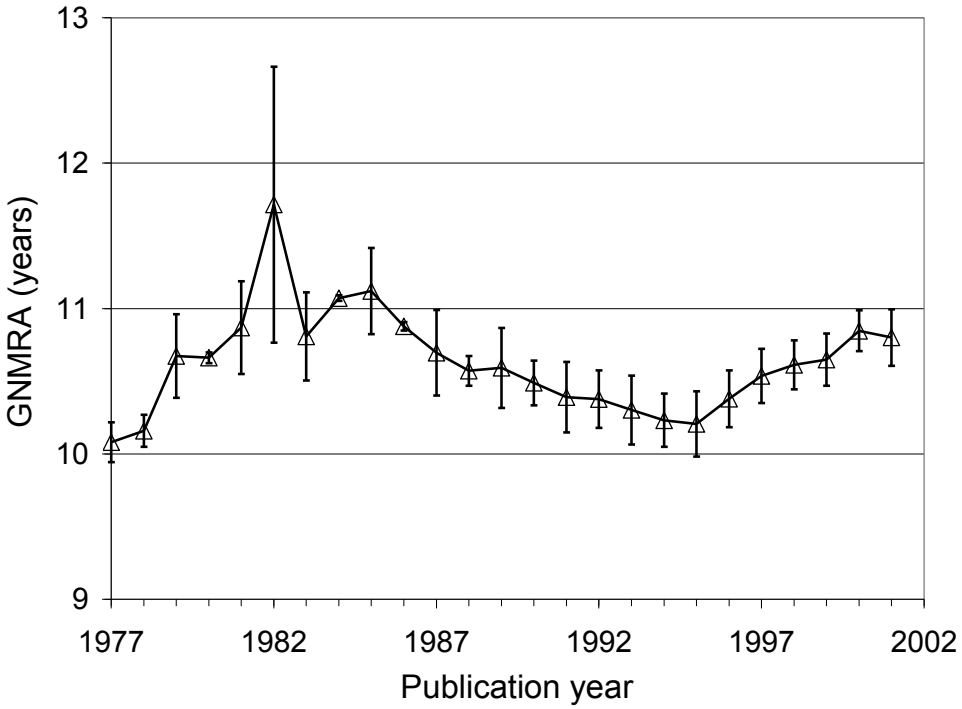


Figure 2: a) top, b) bottom

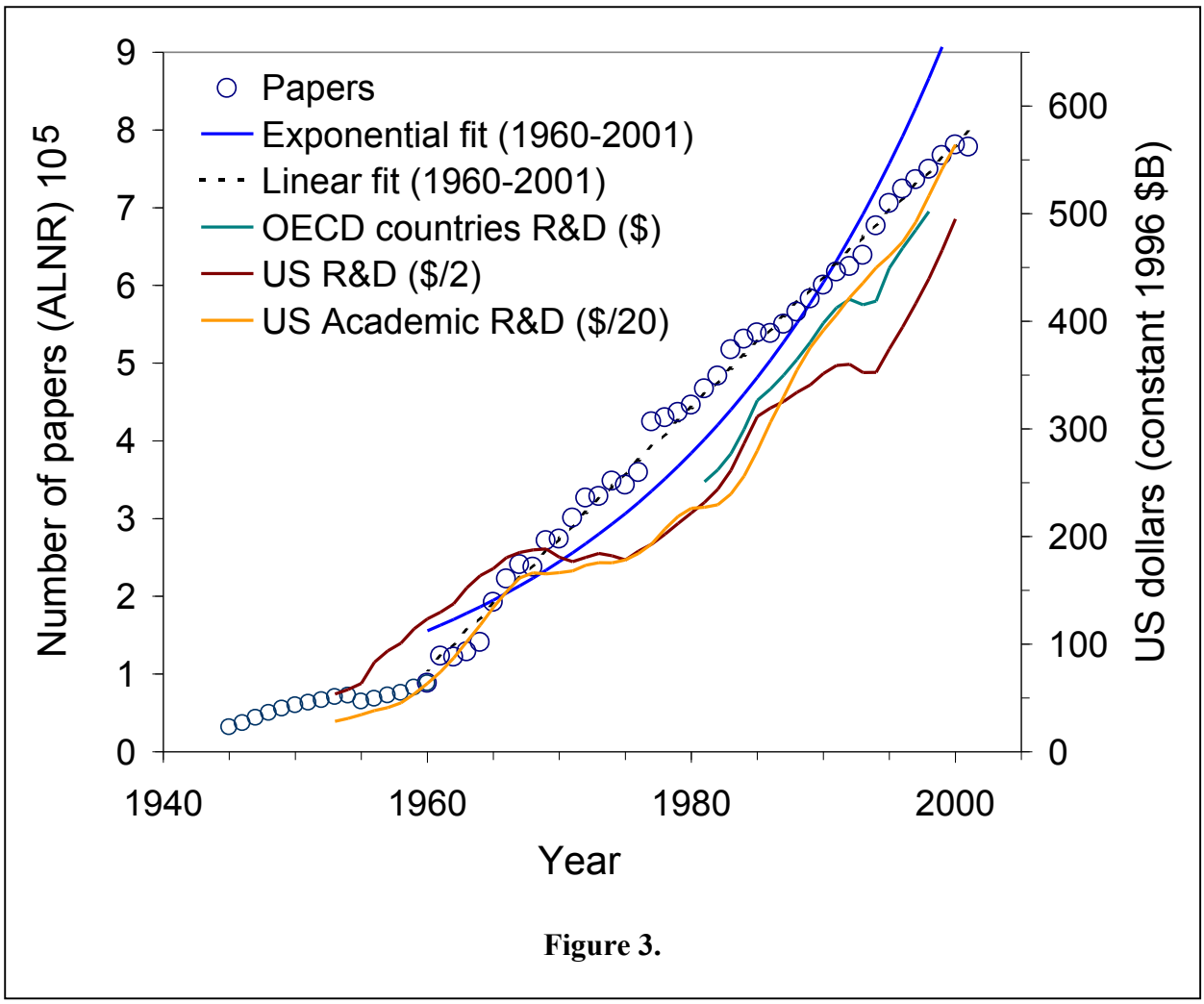

Papers
Exponential fit (1960-2001)
Linear fit (1960-2001)
OECD countries R&D ($)
US R&D ($/2)
US Academic R&D ($/20)
Number of papers (ALNR) 10^5
US dollars (constant 1996 $B)
Year
0
1
2
3
4
5
6
7
8
9
0
100
200
300
400
500
600
1940
1960
1980
2000

Figure 3.

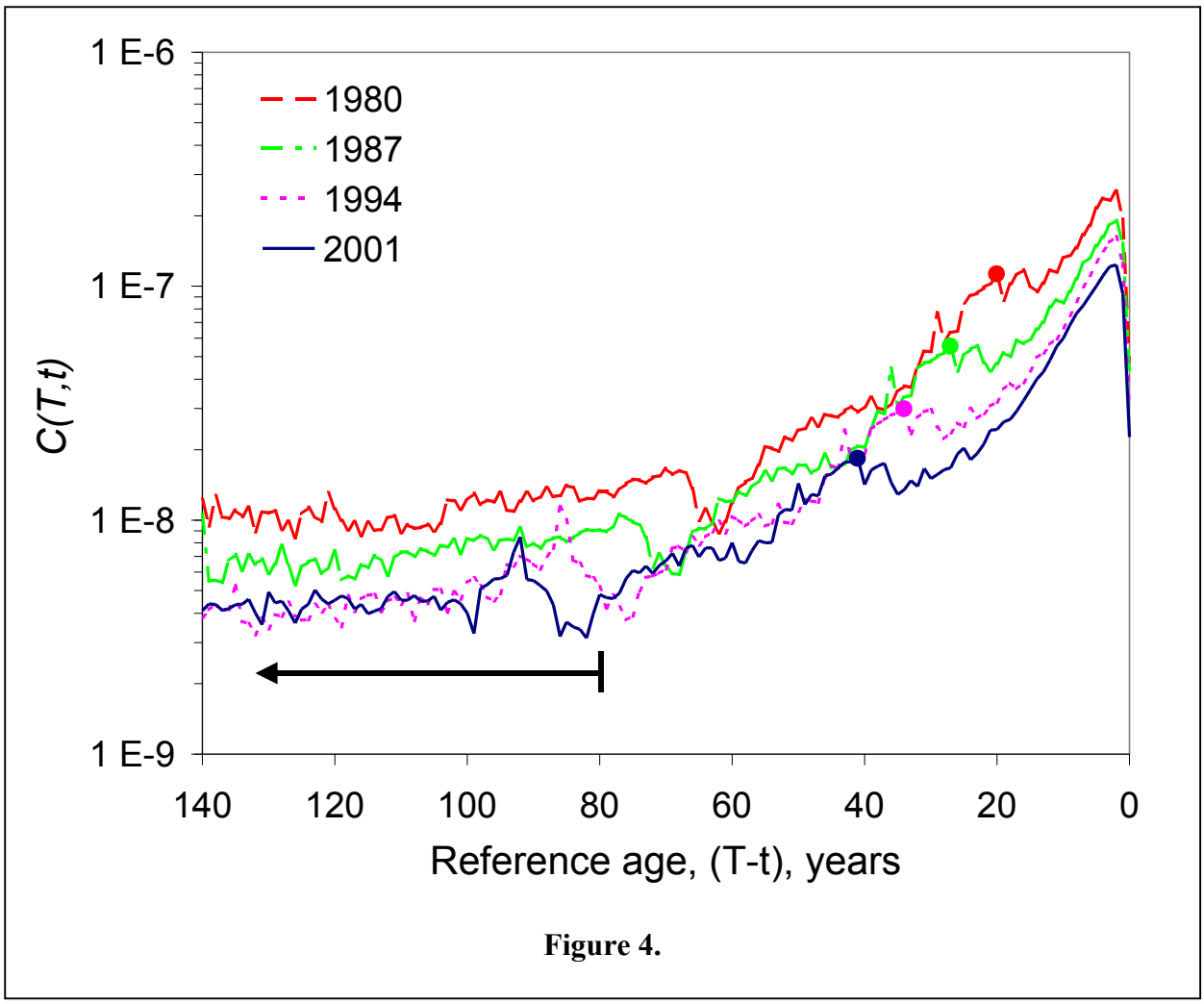

1 E-6
1 E-7
1 E-8
1 E-9
C(T,t)
1980
1987
1994
2001
140
120
100
80
60
40
20
0
Reference age, (T-t), years

Figure 4.