The Rhythm of Science, the Rhythm of *SCIENCE*[1]

Liming Liang[* **], Ronald Rousseau[** ***] and Fei Shi[****]

[*]*pllm@public.xxptt.ha.cn*
Institute for Science Technology and Society, Henan Normal University, Xinxiang, 453007 (China)

[**]University of Antwerp (UA), IBW, Universiteitsplein1, 2600, Wilrijk (Belgium)

[***]ronald.rousseau@khbo.be
KHBO (Association K.U.Leuven), IWT, Zeedijk 101, 8400 Oostende (Belgium)

[****]Thomson Scientific China Office, No 2 Kexueyuan South Road, Beijing 100080 (China)

Abstract
The rhythm of science may be compared to the rhythm of music. The *R* and *T* indicators studied in this article are complex indicators, trying to reflect part of this rhythm. The *R* indicator interweaves publication and citation data over a long period. *T* constructs an input-output relationship in knowledge production. In this way the *R*- and *T*-sequences can be used to describe the evolutionary rhythm of science considered from two different aspects. As an example the R and T sequences of the journal *Science* from 1945 on are calculated.

Introduction
What is science? How does science evolve? These questions refer to two of the most interesting problems in the philosophy and sociology of science. In general, people consider science to consist of the activities of knowledge production as well as the system of knowledge itself (Kuhn, 1962). Scientific documents, in printed or electronic form, are often the final products of knowledge production. They carry scientific information not only to contemporaries, but also to the next generations. Therefore, analyzing publication and citation data of scientific documents is a way of approaching the questions stated in the first sentences. Publication and citation data, being two basic scientometric indicators, have been playing an essential role in the study of science. As such, they reflect, to a large extent, the process of scientific evolution. Like the evolution of living things, the evolution of science has its own rhythm. In the history of western science, we see between the two prosperous periods of the ancient Greece-Rome science and the science of the Renaissance, the less frugal period of the Middle Ages. Then, more recently for four centuries we experienced a series of peaceful interludes punctuated by violent intellectual revolutions (Kuhn, 1962). Focusing on modern science, we can perceive its pulse by studying scientific documents and the time series of their publications and citations. For example, the publication of thousands of articles on superconductivity and the large scale of citing these articles represent a climax in the science of the end of the 1980's and beginning 1990's.

Apart from global and average publication and citation indicators, relative indicators such as the (modified) impact factors (hereafter IF for short) form another class of important scientometric indicators. A time series of IFs can roughly reflect the evolutionary rhythm of a scientific field, country, or journal during a certain period.

The traditional IF, sometimes referred to as the Garfield-Sher IF, has been generalized by information scientists after it was put forward in 1963 (Garfield & Sher, 1963). Whatever version is used, its calculation is always based on observed values of publications and citations. For this reason we call these impact factors observation-based IFs. In this paper we suggest another type of IF, namely expectation-based IFs (see further for their definition). Furthermore, we define and study ratios of observation-based and expectation-based IFs, leading to relative IFs. This relative indicator series has

[1] The work presented in this paper was supported by the National Natural Science Foundation of China by grant no. 70373055.

Proceedings of ISSI 2005, pages 398–405.
Edited by P. Ingwersen and B. Larsen.

been introduced in (Liang, 2005) under the name of *R*-sequence. Meanwhile we have defined many more sequences of this type (Liang, work in preparation). As will be shown in this contribution, sequences of relative IFs yield another view on the scientific evolution of a field, a country, an institute, or a journal. In our opinion, this new view is in general more informative than, e.g. a simple time series of citations, publications or classical, i.e. Garfield-Sher, impact factors. As a case study, we present a relative IF sequence describing the rhythm of a highly visible journal, namely the journal *Science*. As *Science* is a multidisciplinary journal we expect it to be representative for evolutions occurring in the key fields of science.

Besides the *R*-sequence (Liang, 2005) we would like to introduce another type of indicator. It is created based on an input-output analogy. We consider an article's references as inputs to the scientific process leading to the generation of this article. The content of these cited articles can be considered as a gift from contemporaries or earlier generations of scientists (perhaps one could also say that they constitute an 'intelligent loan', because, as a system, science expects rewards from this loan). Uses, codified as citations, of the new article, i.e. the research results published in it, can then be considered as outputs. So, we consider the following process: research results, codified as articles and acknowledged as references, are inputs for a new article. Outputs consist of the citations this new article receives over time. This idea is illustrated in Figure 1. Note that inputs are fixed, but outputs grow dynamically over time. It is this dynamic aspect that will lead us to the new indicator. Instead of one article one may also consider a larger item set, consisting of a related group of articles. Examples of such item sets are: all articles published during a given time period in one particular journal, one institute, or even one country. Of course the item set may also consist of a single article.

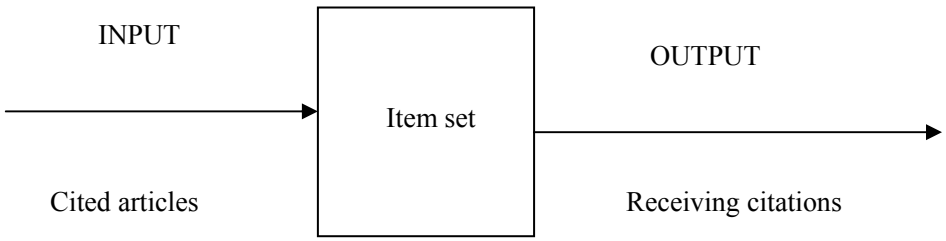

Figure 1: Dynamic input-output process leading to the yield indicator

The time interval, needed to accumulate the same number of citations as the number of references in the item set is defined as the yield period. Articles published in different years may influence scientists in different ways. Therefore the yield period for a journal, institute, or country will fluctuate. Our new indicator (defined more precisely in the next section) can be considered as a kind of time-dependent indicator for the speed with which rewards are reaped from the social-intelligent input of the scientific system into the item set. Note that, generally speaking, a review article requires large inputs, hence the yield period is expected to be larger than that of the average research article. Again, we selected *Science* as a case study.

Methodology

The observation-based IF is defined based on publication and citation data. Suppose we are exploring a journal (it may also be a field, a country, etc.), let the time span be from year 1 to year n, let P_i denote the number of publications in year i, and C_{ij} the number of citations received in the year j by items published in the year i. Then the Garfield-Sher IF of this journal for the year i is defined as

$$IF_i = \frac{C_{i-2,i} + C_{i-1,i}}{P_{i-2} + P_{i-1}}$$

This IF can be generalized by changing the time span of publications, as well as that for citations (Ingwersen et al., 2001; Frandsen & Rousseau, 2005). Thus, the average cited times per paper published in year *i,* is also a kind of observation-based IF, denoted as O_i .

$$O_i = \frac{\sum_{j=i}^{n} C_{ij}}{P_i}$$

We construct the corresponding expectation-based IF, denoted as E_i

$$E_i = \sum_{k=1}^{n-i+1} C_k$$

Here C_k denotes the average number of citations per paper in the *k-th* year after its publication ($k = 1$ to n-i+1, where $k = 1$ refers to the publication year).

Error! Objects cannot be created from editing field codes.

Then, $R_i=O_i/E_i$ is a new indicator, the relative IF for the year *i* with respect to the period *[1,n]* (Liang, 2005). The relative IF has the following properties, distinguishing it from traditional observation-based IFs.

First, the data used in the calculation of an observation-based IF are localized in time. For example, when we calculate the Garfield-Sher IF of the year *t*, only P_{t-1} , P_{t-2} , $C_{t-1,t}$ and $C_{t-2,t}$ are used, other P_i and C_{ij} are not touched. The expectation-based IF, however, is not localized on some fixed years. When we calculate E_i , all the P_i (i=1 , 2 , … , n) join the calculation and all the C_{ij} (i=1 to n, j=1 to n, $j \geq i$) are used in the formula. Therefore, the E_i and hence also the R_i reflect a broader relation among publications and citations in time. Secondly, it is easy to prove that $\sum_{i=1}^{n} E_i = \sum_{i=1}^{n} O_i$, see (Liang, 2005). This equality shows that the essential property of the E_i is a redistribution of all observed citation frequencies over various years in order to reach a theoretical or ideal state. Therefore, the essential meaning of the R_i is to compare observed and ideal states. If in a year the observed value is smaller than the theoretical value, we say that the achievements in that year are lower. Contrarily, the achievements are relative high when the observed value is higher than the expected one. The rhythm expressed by the sequence of R_i – values (here after *R*-sequence for short) is the sequence of relative ups and downs of the item set's scientific achievements and influence.

It is easy to explain the calculation of the yield indicator. Consider a journal as an example. Denote the total number of references cited in its articles in year *i* as L_i . By A_{ih} we denote the cumulative number of citations received by the items published in year *i* since the publication year *i* until year *h*. $A_{ih} = \sum_{j=i}^{h} C_{ij}$. Comparing L_i with all the $A_{ih,}$ we may find a time *t* such that, $A_{it} \leq L_i < A_{i,\,t+1.}$ Then, for publication year *i*, the time interval needed to accumulate a number of citations equal to L_i is T_{1i} , where T_{1i} is calculated as: $T_{1i} = t + (L_i - A_{it}) / (A_{i,\,t+\,1} - A_{it})$. This formula uses linear interpolation, assuming – as an approximation - a uniform distribution of citations over one year. All the T_{1i} (i = 1 to n) form a time series $T_1 = (T_{1i})_i$. The sequence T_1 is another rhythm sequence for this journal, a rhythm sequence of yields.

Similarly, we may define, at least theoretically, a sequence $(S_k)_k$ where each S_k is a time series. Each of its components, S_{ki}, is defined through the following requirements:

$$A_{it} \leq kL_i < A_{i,\,t+1} \text{ and } S_{ki} = t + (kL_i - A_{it}) / (A_{i,\,t+1} - A_{it}).$$

The sequence $S_1 = T_1$. Clearly, S_2, S_3, … can be described as the sequences where the yield (number of citations received) is twice, thrice, … that of the number of references. We refer to S_1, S_2, S_3, … as the first, second, third, etc cumulative yield sequences.
It is also of interest to study the sequences $(T_k)_k$ where

$$T_1 = S_1 \text{ and } T_{ki} = S_{ki} - S_{k-1,i}.$$

Recall that the index i refers to a publication year. T_{ki} denotes the time (expressed in years) required for documents published in the year i, to go from a total of citations equal to $(k-1)L_i$ to a total equal to kL_i. The sequences T_k are called the yield sequences.

A simple theoretical result
Assume that the received citation distribution can be described by a Weibull function (Burrell, 2002, Börner et.al, 2004). Then the cumulative number of citations is given as,

$F(t) = TOT.\left(1 - e^{-\left(\frac{t}{\theta}\right)^{\beta}}\right)$, where TOT is the total number of citations received since publication, θ (> 0) is the scale parameter and β (>0) is the shape parameter. Then, assuming that the number of references is L, S_k is determined through the relation: $F(S_k) = kL$. From this equation we find $TOT\left(1 - e^{-\left(\frac{S_k}{\theta}\right)^{\beta}}\right) = kL$, as long as kL ≤ TOT. A simple calculation then yields:

$$S_k = \theta\left(-\ln\left(\frac{TOT - kL}{TOT}\right)\right)^{1/\beta}.$$

In order to study how S_k depends on k we study the function $y_\beta = (-\ln(1-x))^{1/\beta}$, for $0 \leq x < 1$. For all values of β, y_β passes through the origin and through the point with coordinates $\left(1 - e^{-1}, 1\right)$. Clearly y is always increasing. For $0 < \beta \leq 1$ the curve is convex. This means that the $(T_k)_k$ are increasing: it takes more and more time to reach the next L citations. Note that β = 1 is the case of an exponentially decreasing citation curve. If β > 1, then the curves y_β are concave on the interval $\left[1, 1 - e^{\frac{1-\beta}{\beta}}\right]$, and convex on the interval $\left[1 - e^{\frac{1-\beta}{\beta}}, 1\right]$. In particular, for $x > 1 - e^{-1}$, the function is always convex. This means that in the case β > 1 the T_k are first decreasing and later increasing: first it takes less and less time to reach the next L citations, then it takes more and more time. This corresponds intuitively with a citation curve that increases first and then decreases, a quite natural situation. Note though that this analysis is performed using continuous variables. In reality, depending on the exact values of L, TOT and β it might be that the T_k are always increasing (if the increasing part of the citation curve is too short).

The rhythm of *Science*

All publication and citation data used to study the rhythm of *Science* are retrieved from ISI's Web of Science. We explored the period from 1945 to 2003, a total of 59 years. The journal *Science* published many types of documents: articles, letters, book reviews, editorials, etc. The total number of published documents over this period of 59 years amounts to 103,586. We found that the mix of different document types changed annually. The proportion of 'normal' articles changes from year to year. As articles usually receive much more citations than letters and other documents, we only use 'normal' research articles in our investigation. Thus our sample set consists of 48,828 articles. These articles contain a total of 1,048,423 references, or an average of 21.5 per article. After their publication in *Science* they attracted a total of 4,206,064 citations over the period 1945-2003. Recall that these 1,048,423 references are considered to be an intellectual input of scientists in the journal *Science.* Note also that these 1,048,423 references are of course not different: many articles were used several times. As we had no access to computerized methods (or a host such as DIALOG) it took eight persons more than two months to collect (and re-collect as an accuracy check) all data. Finally, we calculated all the P_i, L_i and C_{ij} (i=1, 2, …, n; $j = i, i$+1,…, n), which are the basic numerical data for the derivation of the rhythm sequences.

SCIENCE' s R-sequences

Based on all the P_i and C_{ij} and using the formulas shown above we calculated O_i, E_i and R_i ($i = 1, 2,$ …, n) for any n, $1 \leq n \leq 59$. Here, the R-sequences of *Science* for n = 30, n = 40, n = 50 and n= 59 are shown (Figure 2).

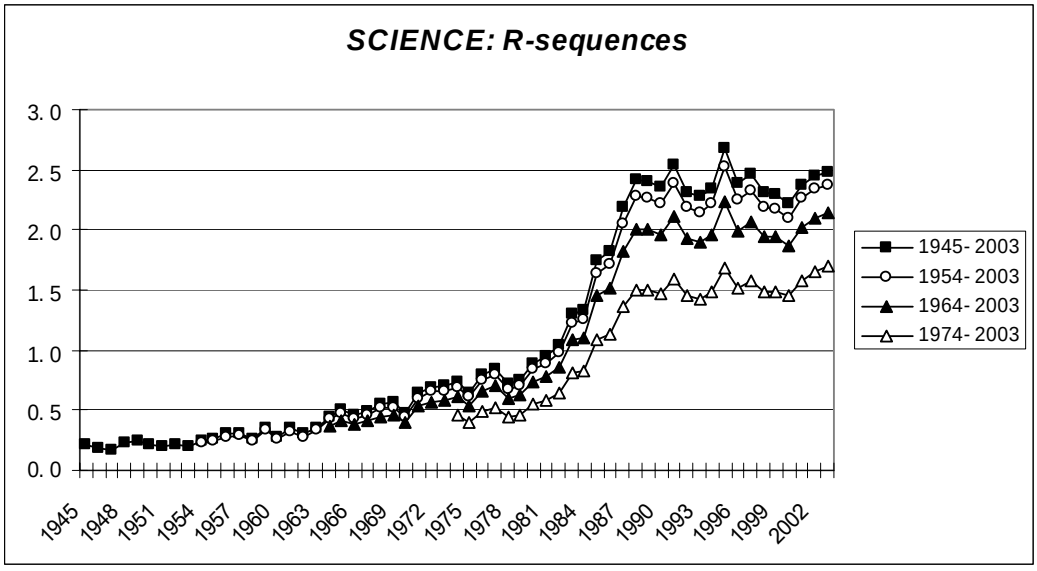

Figure 2: *Science*'s R-sequences

Science, as any other journal, clearly has its own rhythm of evolution. The four curves shown in Figure 2 have a similar shape; *R*-values are only on different levels. Concentrating on the 59 years' curve, and commenting on its changing trend, we note that the evolution of *Science* over the latest 59 years went through three phases. During the period 1945-1981 *Science'*s *R*-values steadily increased from 0.22 to 0.95 with small yearly fluctuations. In 1982 its *R*-value exceeded the value one for the first time. The year 1982 is also the beginning of a six-year period of fast development. During those six years, *Science* moved up to a higher level of academic influence (at least as reflected through the ISI database; this is an important caveat, as the journal content of this database changes!). Its *R*-value becomes larger than two. After 1987 *Science* entered a third phase in its evolution, staying on a high plateau, again with only minor fluctuations. The *R*-sequence forms an *S*-curve, reminding us of Verhulst's logistic curve, as mentioned by Price in "Little Science, Big Science" (Price, 1963).

Based on the collected data we recalculated the Garfield-Sher impact factor for the whole period 1947-2003 (See Appendix). Surprisingly the Pearson correlation coefficient between this series of impact

factors and the *R*-sequence is very high, namely 0.9775. The two curves behave very similarly, except for the latest years. Figure 3 illustrates this. Note that the Garfield-Sher impact factor is on the x-axis, the *R*-value on the y-axis. Because the impact factor increases the x-axis also acts more or less as a time axis.

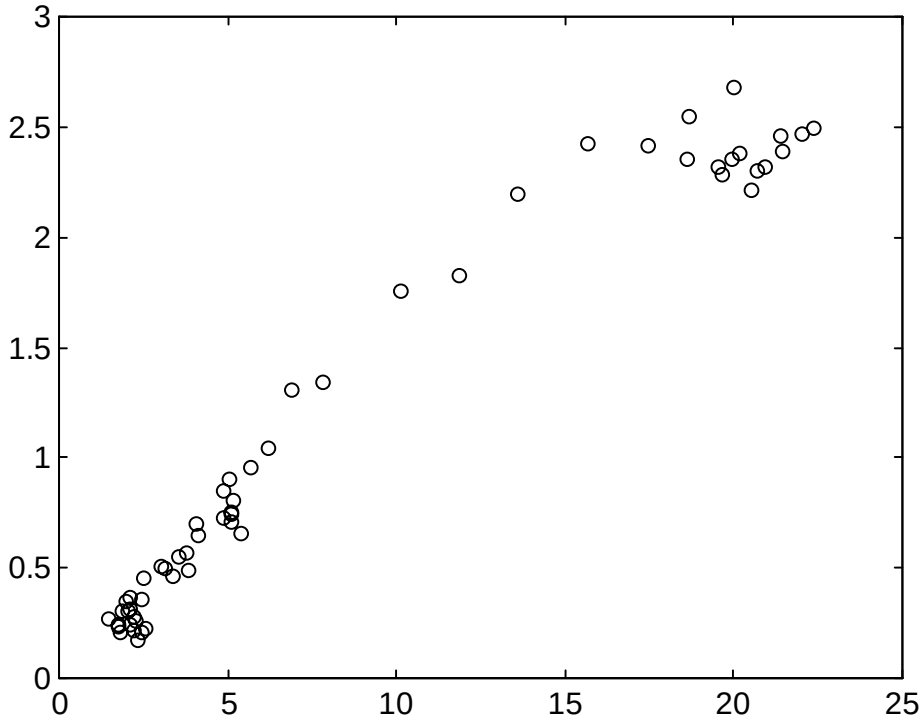

Figure 3: The journal *Science*: Garfield-Sher impact factor versus *R*-value

The Garfield-Sher impact factor only takes two years into account, while the rhythm indicators are based on a much longer period. We explain the remarkable phenomenon of their high correlation by the fact that Science is a top journal: its articles are read and studied almost immediately after publication. After two years the journal issue does not contain any surprises anymore, so that citations received during the first two years (not including the publication year) are almost perfect predictors for citations in the following years. We conjecture that such a correlation does not hold for 'the average journal'. This has been checked for JASIST where indeed the IF-curve and the R-curve behave in quite different ways. Hence, we certainly believe that the *R*-sequence and the sequence of Garfield-Sher impact factors contain different information. We intend to study this phenomenon for more journals taken from different fields.

SCIENCE' s T-sequence

The T-sequence shows us how *Science* draws knowledge from the scientific community on the one hand, and at the same time, contributes knowledge to society by being used (shown through the citations it received) by this scientific community. Figure 4 shows *Science*'s three T-sequences, T_1, T_2 and T_3.

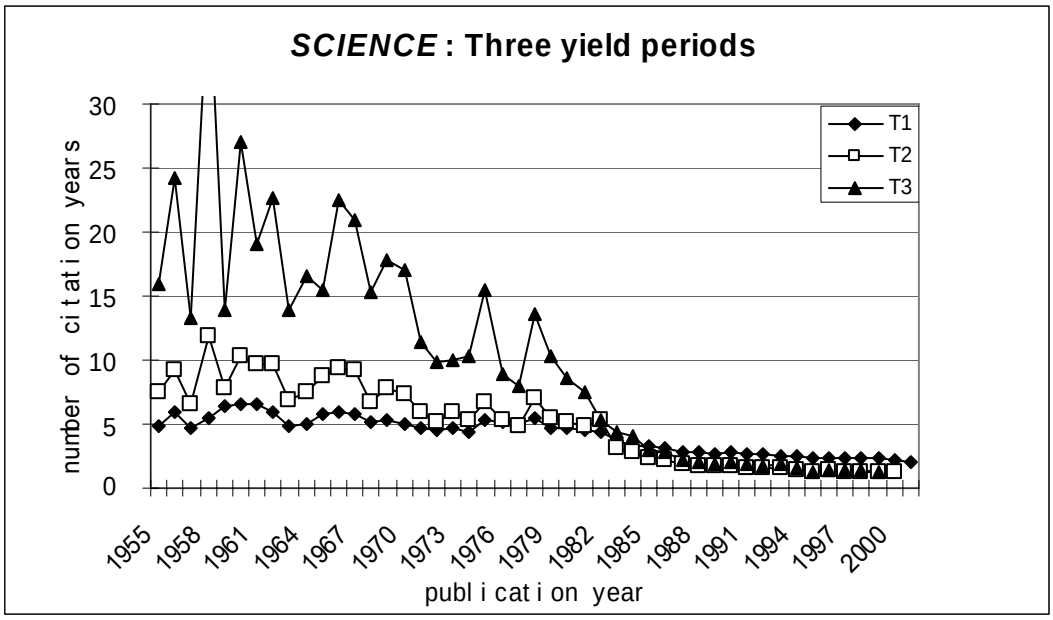

Figure 4: *Science*'s first three yield periods

The three T sequences of *Science* exhibit the same trend: from the earlier time to the present the T-values decrease gradually. That means that all yield periods of *Science* are shortening. During the period 1945-1981 $T_{1i} \leq T_{2i} \leq T_{3i}$, with generally decreasing differences between them. The T_1 curve is the smoothest, T_2 is more fluctuating, and T_3 is the more irregular of the three. After 1981, and certainly after 1985 $T_{2i} \leq T_{1i}$, as well as $T_{3i} \leq T_{1i}$, and though the difference between T_{3i} and T_{2i} gets smaller and smaller, T_{3i} is never smaller than T_{2i}. All this indicates that in this period *Science* accrues citations at a very fast rate. Even the initial period between publication and first citations has become visible ($T_{3i} \leq T_{1i}$). We also note that for all i the T_{1i} , T_{2i} and T_{3i} are larger than 1.

Conclusion and discussion

The rhythm of Science (and of journals such as *Science*) may be compared to the rhythm of a music piece. The *R* and *T* indicators studied in this article are complex indicators, trying to reflect part of this rhythm. The *R* indicator interweaves publication and citation data over a long period. *T* constructs an input-output relationship in knowledge production. In this way the *R*- and *T*-sequences can be used to describe the evolutionary rhythm of science from two different aspects. If it were feasible to obtain the required data, it would become possible to demonstrate how science evolves in a field, a country and even in the world as whole. We are aware though of the limitations of our methodology for measuring the rhythm of science. The main limitation lies in the data collection. Even a database such as the Web of Science can never cover all publications and citations. Therefore, we can never obtain all citation data for an article, or an item set in general. Maybe, in the near future, the evolution of the Internet and its search engines will, however, reduce this limitation.

Finally, we will discuss some questions related to the indicators studied in this article.

Question 1: Considering the *R*-sequence we wonder: how long will *Science*'s third period continue? Does there exist a "ceiling", blocking the further increase of the *R* curve?

Question 2: The ISI database changes dynamically over time. Journals enter and leave the pool. The number of published articles included in the database increases (Jin & Rousseau, 2005). Since 1945 the number of journals itself has increased considerably. Hence the *R*-sequence described in our article is the result of two forces: external ones resulting from the changes in the ISI database and internal ones, reflecting the relative position of *Science* itself. How can these two forces be decomposed? It seems that a normalisation method would be a good start. We are preparing the necessary data in order to perform this normalisation.

Question 3: We saw that the yield period of *Science* gets shorter and shorter, but up to now all T values are larger than one. Is there a minimum yield period? If there is, what is it? To answer this question the citing behaviour of authors, the publication period of the citing and cited journals, and the emergence of more and more electronic journals should be taken into account. In the limit the question becomes: is it possible to receive L citations in one day, one hour, one minute? Yes if L = 1. But what about a more realistic situation?

Acknowledgement

The authors appreciate very much all the students who worked very hard to search and check the data. They are: Ma Xiaohua, Li Puyu, Zhang Lin, Zhong Zhen, Yang Weixue, Xu Chao Feng, Zhang Li and Liu Qiuge. They also thank the anonymous ISSI reviewers for helpful suggestions.

References

Börner, K., Maru, J.T. and Goldstone, R.L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101, suppl.1, 5266-5273.

Burrell, Q.L. (2002). Modelling citation age data; simple graphical methods from reliability theory. *Scientometrics*, 55, 273-285.

Frandsen, T.F. and Rousseau, R. (2005). Article impact calculated over arbitrary periods. *Journal of the American Society for Information Science and Technology*, 56, 58-62.

Garfield, E. and Sher, I.H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, 14, 195-201.

Ingwersen, P., Larsen, B., Rousseau, R. and Russell, J. (2001). The publication-citation matrix and its derived quantities. *Chinese Science Bulletin*, 2001, 46(6), 524-528.

Jin, B. and Rousseau, R. (2005). China's quantitative expansion phase: exponential growth but low impact. *These Proceedings*.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Liang, L. (2005). The R-sequence: a relative indicator for the rhythm of science. *Journal of the American Society for Information Science and Technology*, 56 (to appear).

Price, D. de Solla (1963). *Little science, big science*. New York: Columbia University Press.

Appendix

Table 1 Recalculated Garfield –Sher impact factors (only articles) of *Science*

Year	IF	Year	IF	Year	IF	Year	IF
1947	2.39	1962	1.93	1977	4.89	1992	19.55
1948	1.79	1963	2.13	1978	4.93	1993	19.69
1949	2.11	1964	2.56	1979	5.12	1994	19.96
1950	2.58	1965	3.04	1980	5.10	1995	20.03
1951	2.49	1966	3.38	1981	5.71	1996	21.50
1952	2.27	1967	3.20	1982	6.25	1997	22.04
1953	1.82	1968	3.58	1983	6.94	1998	20.96
1954	1.78	1969	3.80	1984	7.84	1999	20.73
1955	1.48	1970	3.89	1985	10.19	2000	20.53
1956	2.05	1971	4.18	1986	11.90	2001	20.21
1957	2.16	1972	4.11	1987	13.63	2002	21.43
1958	2.29	1973	5.11	1988	15.72	2003	22.43
1959	2.49	1974	5.16	1989	17.51		
1960	2.24	1975	5.40	1990	18.63		
1961	2.02	1976	5.19	1991	18.73		