Scientometrics
DOI 10.1007/s11192-009-0099-6

# Probing the *h*-core: an investigation of the tail–core ratio for rank distributions

**Fred Y. Ye · Ronald Rousseau**

Received: 9 August 2009


**Abstract** The set of citations received by a set of publications consists of citations received by articles in the *h*-core and citations received by articles in the *h*-tail. Denoting the cardinalities of these fours sets as $C$, $P$, $C_H$ and $C_T$ we introduce the tail-core ratio ($C_T/C_H$) and show that in practical cases this ratio tends to increase. Introducing further the *k*-index, defined as $k = (C/P)/(C_T/C_H)$, we show that this index decreases in most practical cases. A power law model is in accordance with these practical observations.

**Keywords** *h*-Index · *h*-Core · *h*-Tail · Tail–core ratio · *v*-Index · *k*-Index

## Introduction

The *h*-index (Hirsch 2005) divides a set of articles into two groups: the first *h* ones, each having at least *h* citations during the period under study and the other ones, each having at most *h* citations. The first set is known as the *h*-core (Rousseau 2006), while the second one has not received a name yet. We propose to call it the *h*-tail, referring to the long tail (sequence of ranks) of articles that do not belong to the *h*-core. In order to designate articles with exactly *h* citations to the core or the tail a secondary criterion, e.g. age, is necessary. However, this delineation plays no role in our article. Note that for rank

F. Y. Ye
Department of Information Resources Management, Zhejiang University, Hangzhou, China
e-mail: yye@zju.edu.cn

F. Y. Ye
Institute of Scientific and Technical Information of China, Beijing, China

R. Rousseau (✉)
Department of Industrial Sciences and Technology, KHBO (Association K.U.Leuven),
Zeedijk 101, 8400 Oostende, Belgium
e-mail: ronald.rousseau@khbo.be

R. Rousseau
Department of Mathematics, K.U.Leuven, Celestijnenlaan 200B, 3001 Heverlee, Leuven, Belgium



Published online: 25 September 2009

distributions the tail consists of sources producing the least number of items, contrary to the case of frequency distributions. As a first step we will try to find out the relation between the *h*-core and the *h*-tail. Which one is usually the largest in terms of publications and in terms of citations? How does the ratio of these numbers, denoted as $C_T/C_H$, change over time? We collected practical data leading to some interesting conclusions.

## Calculation method

We recall that the idea of an *h*-index and an *h*-core can be applied to many source–item relations (Schubert and Glänzel 2007; Ye and Rousseau 2008; Egghe 2010) and over many time windows (Liang and Rousseau 2009). In general we assume that there are $P$ sources and $C$ citations. By definition the *h*-core consists of $h$ sources, hence there are $P$–$h$ sources in the *h*-tail. The ratio $h/P$ is called the *v*-index (Riikonen and Vihinen 2008). The number of citations in the *h*-core, denoted as $C_H$, has no upper limit but is at least equal to $h^2$. The number of citations in the *h*-tail, denoted as $C_T$, is at most $(P–h)h$ and has zero as a lower limit.

The square root of $C_H$ was introduced in (Jin et al. 2007) as a complementary *h*-type index denoted as $R$:

$$R = \sqrt{C_H} \tag{1}$$

The value $C_H - h^2 = e^2$ was recently introduced as another complementary *h*-type index (Zhang 2009). Here the symbol $e$ stands for excess citations. It is claimed that $e$ and $h$ are independent. Although the term independent is not precisely defined in Zhang (2009), we would like to point out that $e$ is a value that can only be calculated if $h$ is given. Given a set of publications and citations it is impossible to calculate the *e*-index directly. One must first determine the *h*-index before the value of the *e*-index can be derived. The relation between the *h*-tail citations and the *h*-core citations is illustrated for a continuous model in Fig. 1. The *h*-core citations consist of the citations in the union of the *h*-area and the *e*-area.

The value $C/P$ represents the average number of citations per publication and is usually called the impact, or the impact factor (in the case of a journal and the well-known special publication–citation window). The impact is a real-valued number with zero as its lower bound and with no theoretical upper bound.

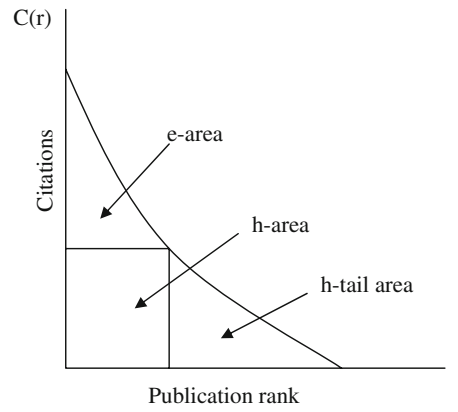


**Fig. 1** A rank-citation curve (in continuous form) subdivided into three areas

## Time series in citation analysis

There are many different time series of interest in citation analysis. In this article we consider a type 2 time series as illustrated in (Liu and Rousseau 2008). The term 'Type 2' means that, for each year Y, we consider the cumulative number of publications from a starting year until year Y and the corresponding cumulative number of citations received by these publications during the period from publication to an end date (the same date for each element in the time series). For our practical examples (see further) this end date is January 1, 2009. When $P(t)$ denotes the cumulative number of publications of a source over a time period with length $t$ and $C(t)$ denotes the corresponding cumulative number of citations received, then often the impact $C(t)/P(t)$ decreases over time as the oldest articles have the longest period to garner citations. For simplicity we will write 'increase' in what follows, when we actually mean 'does not decrease'.

In a type 2 time series analysis $h(t)$ increases by definition. Consequently also $C_H(t)$ and $C_T(t)$ increase. We first use some fictitious examples in order to show that the impact as defined above can (strictly) decrease, (strictly) increase, or stay unchanged.

The first three examples of Table 1 have an *h*-index that does not change from year 1 to year 2; the last case has an *h*-index that increases. In each case only one new article is published in year 2. These examples illustrate that the impact C/P can strictly increase, strictly decrease or be unchanged and that $C_T$ can either strictly increase or be unchanged.

We are now interested in $C_T(t)/C_H(t) = (C(t) - C_H(t))/C_H(t)$, called the tail–core ratio as a function of time, and want to compare the impact to this tail–core ratio. For this reason we introduce the *k*-index, also a time-dependent index, defined as:

$$k(t) = \frac{C(t)}{P(t)} \Big/ \frac{C_T(t)}{C_H(t)} = \frac{C(t)C_H(t)}{P(t)(C(t) - C_H(t))} \tag{2}$$

The tail–core ratios and *k*-values for the three cases of Table 1 are given in Table 2.

**Table 1** Four cases

| Years | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| 1a. Data | | | | |
| 1 | Art. 1-4 cit | Art. 1-3 cit | Art. 1-3 cit | Art. 1-3 cit |
| | Art. 2-1 cit | Art. 2-2 cit | Art. 2-2 cit | Art. 2-3 cit |
| | | Art. 3-1 cit | Art. 3-1 cit | Art. 3-1 cit |
| 2 | Art. 1-4 cit | Art. 1-3 cit | Art. 1-7 cit | Art. 1-5 cit |
| | Art. 2-1 cit | Art. 2-2 cit | Art. 2-3 cit | Art. 2-3 cit |
| | Art. 3-0 cit | Art. 3-2 cit | Art. 3-2 cit | Art. 3-3 cit |
| | | Art. 4-1 cit | Art. 4-1 cit | Art. 4-1 cit |
| 1b. Indices | | | | |
| 1 | $C/P = 5/2$ | $C/P = 2$ | $C/P = 2$ | $C/P = 7/3$ |
| | $h = 1$ | $h = 2$ | $h = 2$ | $h = 2$ |
| | $C_T = 1$ | $C_T = 1$ | $C_T = 1$ | $C_T = 1$ |
| 2 | $C/P = 5/3$ | $C/P = 2$ | $C/P = 13/4$ | $C/P = 3$ |
| | $h = 1$ | $h = 2$ | $h = 2$ | $h = 3$ |
| | $C_T = 1$ | $C_T = 3$ | $C_T = 3$ | $C_T = 1$ |

**Table 2** Tail–core ratios and *k*-index for the four cases shown in Table 1

| Years | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| 1 | $C_T/C_H = 1/4$<br>$k = 10$ | $C_T/C_H = 1/5$<br>$k = 10$ | $C_T/C_H = 1/5$<br>$k = 10$ | $C_T/C_H = 1/6$<br>$k = 14$ |
| 2 | $C_T/C_H = 1/4$<br>$k = 20/3$ | $C_T/C_H = 3/5$<br>$k = 10/3$ | $C_T/C_H = 3/10$<br>$k = 13/2$ | $C_T/C_H = 1/11$<br>$k = 33$ |

**Table 3** A journal: *Scientometrics*

| Publication period | $P$ | $C$ | $C_H = R^2$ | $C_T$ | $C/P$ | $h$ | $v = h/P$ |
|---|---|---|---|---|---|---|---|
| 98–98 | 102 | 863 | 410 | 453 | 8.46 | 16 | 0.157 |
| 98–99 | 228 | 1,389 | 515 | 874 | 6.09 | 19 | 0.083 |
| 98–00 | 339 | 2,214 | 639 | 1,575 | 6.53 | 22 | 0.065 |
| 98–01 | 440 | 3,026 | 852 | 2,174 | 6.88 | 24 | 0.055 |
| 98–02 | 541 | 3,934 | 1,008 | 2,926 | 7.27 | 26 | 0.048 |
| 98–03 | 635 | 4,615 | 1,059 | 3,556 | 7.27 | 27 | 0.043 |
| 98–04 | 736 | 5,368 | 1,172 | 4,196 | 7.29 | 28 | 0.038 |
| 98–05 | 850 | 5,913 | 1,256 | 4,657 | 6.96 | 29 | 0.034 |
| 98–06 | 996 | 6,489 | 1,384 | 5,105 | 6.52 | 31 | 0.031 |
| 98–07 | 1,130 | 6,686 | 1,384 | 5,302 | 5.92 | 31 | 0.027 |
| 98–08 | 1,262 | 6,730 | 1,384 | 5,346 | 5.33 | 31 | 0.025 |

**Table 4** An institution: the University of Heidelberg

| Publication period | $P$ | $C$ | $C_H = R^2$ | $C_T$ | $C/P$ | $h$ | $v = h/P$ |
|---|---|---|---|---|---|---|---|
| 98–98 | 2,676 | 57,972 | 19,920 | 38,052 | 21.66 | 98 | 0.037 |
| 98–99 | 5,240 | 106,594 | 26,883 | 79,711 | 20.34 | 122 | 0.023 |
| 98–00 | 7,672 | 154,874 | 35,390 | 119,484 | 20.19 | 136 | 0.018 |
| 98–01 | 10,210 | 200,974 | 40,225 | 160,749 | 19.68 | 141 | 0.014 |
| 98–02 | 12,651 | 246,564 | 47,062 | 199,502 | 19.49 | 150 | 0.012 |
| 98–03 | 15,530 | 287,828 | 49,888 | 237,940 | 18.53 | 157 | 0.010 |
| 98–04 | 18,327 | 325,927 | 56,015 | 269,912 | 17.78 | 161 | 0.009 |
| 98–05 | 21,772 | 361,417 | 58,355 | 303,062 | 16.60 | 166 | 0.007 |
| 98–06 | 25,073 | 387,121 | 61,754 | 325,367 | 15.44 | 169 | 0.007 |
| 98–07 | 28,654 | 398,791 | 61,754 | 337,037 | 13.92 | 169 | 0.006 |
| 98–08 | 32,605 | 401,676 | 61,754 | 339,922 | 12.32 | 169 | 0.005 |

The tail–core ratio increases for cases 2 and 3, is unchanged for case 1 and decreases strictly for case 4. The *k*-index decreases strictly for cases 1, 2 and 3, while it increases strictly for case 4. These examples show that, at least theoretically there is no fixed relation

**Table 5** Patent data: Motorola

| Period | $P$ | $C$ | $C_H = R^2$ | $C_T$ | $C/P$ | $h$ | $v = h/P$ |
|---|---|---|---|---|---|---|---|
| 98–98 | 1,648 | 22,084 | 4,699 | 17,385 | 13.40 | 57 | 0.035 |
| 98–99 | 2,979 | 38,358 | 6,117 | 32,241 | 12.88 | 66 | 0.022 |
| 98–00 | 4,196 | 52,489 | 7,056 | 45,433 | 12.51 | 71 | 0.017 |
| 98–01 | 5,317 | 61,873 | 7,323 | 54,550 | 11.64 | 73 | 0.014 |
| 98–02 | 6,502 | 71,036 | 7,507 | 63,529 | 10.93 | 75 | 0.011 |
| 98–03 | 8,065 | 78,381 | 7,609 | 70,772 | 9.72 | 75 | 0.009 |
| 98–04 | 9,410 | 82,768 | 7,609 | 75,159 | 8.80 | 75 | 0.008 |
| 98–05 | 10,403 | 83,734 | 7,609 | 76,125 | 8.05 | 75 | 0.007 |
| 98–06 | 11,216 | 84,060 | 7,609 | 76,451 | 7.49 | 75 | 0.007 |
| 98–07 | 12,273 | 84,135 | 7,609 | 76,526 | 6.86 | 75 | 0.006 |
| 98–08 | 13,560 | 84,142 | 7,609 | 76,533 | 6.21 | 75 | 0.006 |

between these indices. Next we will investigate the tail–core ratio and the *k*-index for some real examples.

## Data

Publication and citation data related to the period 1998–2008 were collected from Thomson Reuters/ISI Web of Science on January 1, 2009, while patent data were collected from the Derwent Innovations Index (DII). Tables 3, 4 and 5 show a representative example for the category of journals, institutions and patent assignees. The letter $P$ stands for number of publications in the case of *Scientometrics* and the University of Heidelberg, while it stands for number of patents in the case of Motorola. It is clear that for this type of time series the *v*-index is expected to decrease, which it does.

The main parameters are illustrated in Figs. 2, 3 and 4.

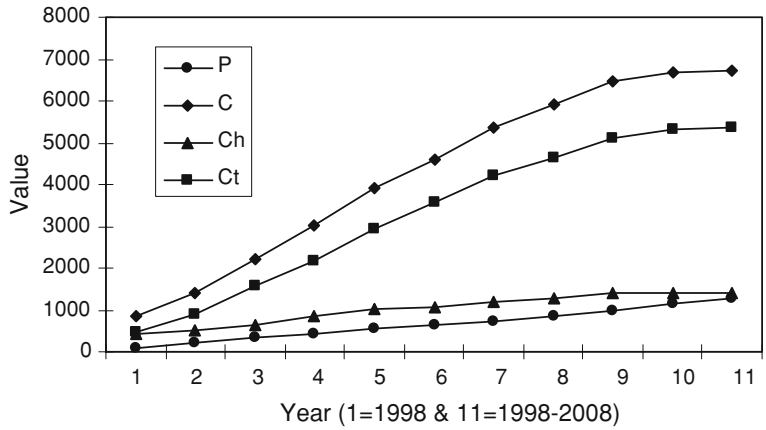


**Fig. 2** Journal citation–publication curves: Scientometrics (1998–2008)

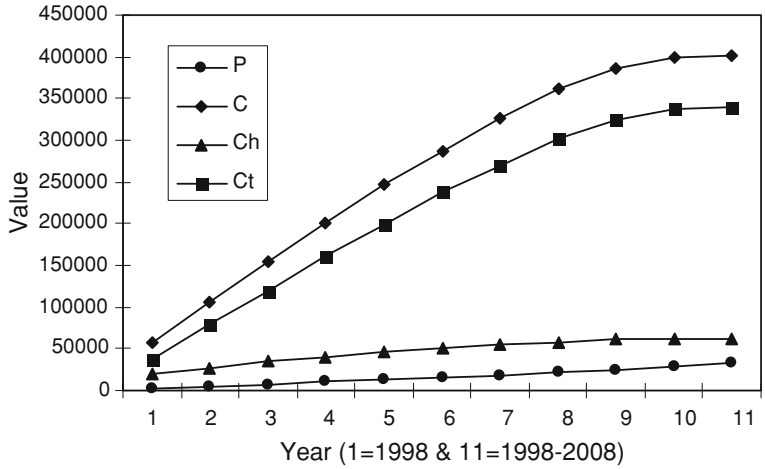


**Fig. 3** Institution citation–publication curves: University of Heidelberg (1998–2008)

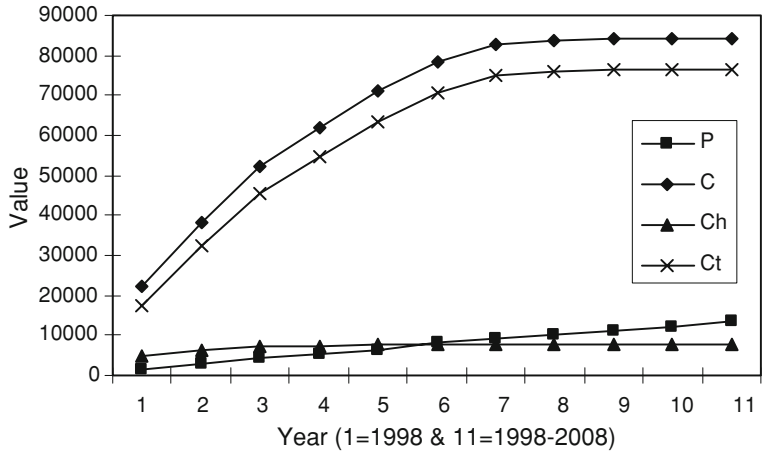


**Fig. 4** Assignee citation–publication curves: Motorola (1998–2008)

## Results and analysis

Based on the data shown in Tables 3, 4 and 5, we calculate the percentage of citations in $C_H$ and in $C_T$ with respect to $C$, the tail–core ratio and the $k$-index. Tables 6, 7 and 8 show the results; for clarity they also include the $h$-index.

Tables 6, 7 and 8 show that in practical situations the percentage of citations in $C_T$ increases, and consequently the tail–core ratio ($C_T/C_H$) increases too. The combined effect of an increasing tail–core ratio and a decreasing impact leads to a decreasing value of the $k$-index. During the same period the $h$-index increases.

**Table 6** Scientometrics

| Publication period | $C_H$ (%) | $C_T$ (%) | $C_T/C_H$ | $k$ | $h$ |
|---|---|---|---|---|---|
| 98–98 | 47.51 | 52.49 | 1.1 | 7.66 | 16 |
| 98–99 | 37.08 | 62.92 | 1.7 | 3.59 | 19 |
| 98–00 | 28.86 | 71.14 | 2.5 | 2.65 | 22 |
| 98–01 | 28.16 | 71.84 | 2.6 | 2.70 | 24 |
| 98–02 | 25.62 | 74.38 | 2.9 | 2.51 | 26 |
| 98–03 | 22.95 | 77.05 | 3.4 | 2.16 | 27 |
| 98–04 | 21.83 | 78.17 | 3.6 | 2.04 | 28 |
| 98–05 | 21.24 | 78.76 | 3.7 | 1.88 | 29 |
| 98–06 | 21.33 | 78.67 | 3.7 | 1.76 | 31 |
| 98–07 | 20.70 | 79.30 | 3.8 | 1.54 | 31 |
| 98–08 | 20.56 | 79.44 | 3.9 | 1.38 | 31 |

**Table 7** University of Heidelberg

| Publication period | $C_H$ (%) | $C_T$ (%) | $C_T/C_H$ | $k$ | $h$ |
|---|---|---|---|---|---|
| 98–98 | 34.36 | 65.64 | 1.9 | 11.34 | 98 |
| 98–99 | 25.22 | 74.78 | 2.97 | 6.86 | 122 |
| 98–00 | 22.85 | 77.15 | 3.38 | 5.98 | 136 |
| 98–01 | 20.02 | 79.99 | 4.00 | 4.93 | 141 |
| 98–02 | 19.09 | 80.91 | 4.24 | 4.60 | 150 |
| 98–03 | 17.33 | 82.67 | 4.77 | 3.89 | 157 |
| 98–04 | 17.19 | 82.81 | 4.82 | 3.69 | 161 |
| 98–05 | 16.15 | 83.85 | 5.19 | 3.20 | 166 |
| 98–06 | 15.95 | 84.05 | 5.27 | 2.93 | 169 |
| 98–07 | 15.49 | 84.51 | 5.46 | 2.55 | 169 |
| 98–08 | 15.37 | 84.63 | 5.50 | 2.24 | 169 |

**Table 8** Motorola

| Patenting period | $C_H$ (%) | $C_T$ (%) | $C_T/C_H$ | $k$ | $h$ |
|---|---|---|---|---|---|
| 98–98 | 21.28 | 78.72 | 3.70 | 3.62 | 57 |
| 98–99 | 15.95 | 84.05 | 5.27 | 2.44 | 66 |
| 98–00 | 13.44 | 86.56 | 6.44 | 1.94 | 71 |
| 98–01 | 11.84 | 88.16 | 7.45 | 1.56 | 73 |
| 98–02 | 10.57 | 89.43 | 8.46 | 1.29 | 75 |
| 98–03 | 9.71 | 90.29 | 9.30 | 1.04 | 75 |
| 98–04 | 9.19 | 90.81 | 9.88 | 0.89 | 75 |
| 98–05 | 9.09 | 90.91 | 10.01 | 0.80 | 75 |
| 98–06 | 9.05 | 90.95 | 10.05 | 0.75 | 75 |
| 98–07 | 9.04 | 90.96 | 10.06 | 0.68 | 75 |
| 98–08 | 9.04 | 90.96 | 10.06 | 0.62 | 75 |

## A power law model

The power law model (Egghe 2005) assumes that the number of sources producing $x$ items is given by the function $F$

$$F : [1, +\infty[ \rightarrow ]0, C] : x \rightarrow \frac{C}{x^{\alpha}} \tag{3}$$

If $\alpha > 2$ then it can be shown (Egghe 2005, p. 115) that the impact is $\frac{\alpha-1}{\alpha-2}$, or $C = \frac{\alpha-1}{\alpha-2}P$. As $h = P^{1/\alpha}$ (Egghe and Rousseau 2006) and A, the average number of citations in the $h$-core, is equal to $\frac{\alpha-1}{\alpha-2}h$ (Jin et al. 2007) it follows that $C_{\mathrm{H}} = \frac{\alpha-1}{\alpha-2}P^{2/\alpha}$ and hence $C_{\mathrm{T}} = \frac{\alpha-1}{\alpha-2}\left(P - P^{2/\alpha}\right)$. Consequently, it follows that in the power law model the tail–core ratio is $\frac{C_{\mathrm{T}}}{C_{\mathrm{H}}} = \frac{P-P^{2/\alpha}}{P^{2/\alpha}} = P^{\frac{\alpha-2}{\alpha}} - 1$. Considering the above quantities as a function of time: $P(t)$, $C(t)$, $\alpha(t)$, $C_{\mathrm{H}}(t)$, $C_{\mathrm{T}}(t)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{C_{\mathrm{T}}(t)}{C_{\mathrm{H}}(t)}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(P^{\frac{\alpha-2}{\alpha}} - 1\right) = \left(P^{\frac{\alpha-2}{\alpha}}\right)\left(\frac{2\alpha'}{\alpha^2}\ln(P) + \frac{\alpha-2}{\alpha}\frac{P'}{P}\right),$$

where all derivatives are taken with respect to time $t$. If we assume now that the impact decreases strictly (as expected for a type 2 time series), then $\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\alpha-1}{\alpha-2}\right) < 0$ or $\frac{\alpha'(\alpha-2)-(\alpha-1)\alpha'}{(\alpha-2)^2} = \frac{-\alpha'}{(\alpha-2)^2} < 0$. Hence $\alpha' > 0$. As also $P' > 0$ for a type 2 time series it follows that $\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{C_{\mathrm{T}}(t)}{C_{\mathrm{H}}(t)}\right) > 0$. Hence the tail–core ratio increases in the power law model. Now $k = \frac{C}{P}/\frac{C_{\mathrm{T}}}{C_{\mathrm{H}}} = \frac{\alpha-1}{\alpha-2}\frac{P^{2/\alpha}}{P-P^{2/\alpha}}$. As the first factor in this last expression as well as the second one is decreasing, $k(t)$ is decreasing in time.

We conclude that the results of the power law model correspond with the observations. Note that, as in other cases, we do not claim that actual sources follow a power law. Indeed, we know that a power law assumption for a type 2 time series leads to problems with practical data (Ye and Rousseau 2008). We just apply this model as a first approximation of an observed frequency distribution, as we did in Jin et al. (2007).

## Conclusion

In this article we have introduced the tail–core ratio in the Hirsch index framework. The ratio of impact over tail–core ratio is called the $k$-index. We have shown that for a type 2 time series the tail–core ratio usually increases in real situations, while the $k$-index decreases. It is shown that a power law model is in accordance with these practical observations.

**Acknowledgements** Fred Y. Ye's work is supported by a grant from the National Natural Science Foundation of China (NSFC Grant No. 70773101), while Ronald Rousseau's work is supported by the National Natural Science Foundation of China Grant No. 70673019. The authors thank graduate student Jianhui Tang for assistance in data collection and Leo Egghe for checking the power law calculations.

## References

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Amsterdam: Elsevier.

Egghe, L. (2010). The Hirsch-index and related impact measures. *Annual Review of Information Science and Technology* (to appear).

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics, 69*(1), 121–129.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572.

Jin, B. H., Liang, L. M., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin, 52*(6), 855–863.

Liang, L. M., & Rousseau, R. (2009). A general approach to citation analysis and an *h*-index based on the standard impact factor framework. In B. Larsen & J. Leta (Eds.), *ISSI 2009; 12th international conference on scientometrics and informetrics* (pp. 143–153). Rio de Janeiro: BIREME and University of Rio de Janeiro.

Liu, Y. X., & Rousseau, R. (2008). Definitions of time series in citation analysis with special attention to the h-index. *Journal of Informetrics, 2*(3), 202–210.

Riikonen, P., & Vihinen, M. (2008). National research contributions: A case study on Finnish biomedical research. *Scientometrics, 77*(2), 207–222.

Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus*, 1(4), 23–25 (in Chinese). English version available at: E-LIS: code 6376.

Schubert, A., & Glänzel, W. (2007). A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics, 1*(2), 179–184.

Ye, F. Y., & Rousseau, R. (2008). The power law model and total career h-index sequences. *Journal of Informetrics, 2*(4), 288–297.

Zhang, C. T. (2009). The e-index, complementing the h-index for excess citations. *PLoS One, 4*(5), e5429.