

A new methodology for ranking scientific institutions

J.F. Molinari¹, A. Molinari^{2*}

¹ Laboratory of Mechanics and Technology, Ecole Normale Supérieure de Cachan, Paris 6,
France

² Laboratory of Physics and Mechanics of Materials, Université Paul Verlaine, Metz, France

* Corresponding author: molinari@lpmm.univ-metz.fr

Abstract:

We extend the pioneering work of J.E. Hirsch, the inventor of the h-index, by proposing a simple and seemingly robust approach for comparing the scientific productivity and visibility of institutions. Our main findings are that i) while the h-index is a sensible criterion for comparing scientists within a given field, it does not directly extend to rank institutions of disparate sizes and journals, ii) however, the h-index, which always increases with paper population, has an universal growth rate for large numbers of papers; iii) thus the h-index of a large population of papers can be decomposed into the product of an impact index and a factor depending on the population size, iv) as a complement to the h-index, this new impact index provides an interesting way to compare the scientific production of institutions (universities, laboratories or journals).

Introduction

The evaluation of the productivity of individual scientists has traditionally relied on the opinion of experts. With the venue of the Internet and access to large databases, including the Web of Science [1], comparing the impact of scientific contributions has gained in automation and simplicity. Recently, a novel and simple-to-use index was proposed by J.E. Hirsch [2]. This index, referred to as h or h-index, is intended to replace the traditional method of paper counting, which advantages scientists who produce a large quantity of weakly cited papers, as well as the total citation count, which may put too much emphasis on few highly cited papers. The h-index is defined as the number h of papers with citation number higher or equal to h [2]. h is thus an indication of both quantity and quality. The h-index can only increase in a scientist's career, and therefore one should compare scientists of a given age group (or compare the slope of h when plotted as a function of time). This novel indicator has gained fast and broad recognition [3,4,5,6], and it is natural to try to extend its use for ranking institutions (be they universities, institutes, laboratories, scientific journals etc...), as was suggested in the original paper [2]. However, the h-index is a function of the number of papers considered and thus large research institutions producing a great number of papers may have a larger h-index than smaller institutions. Thus, simply considering h may not be a fair way to evaluate the intrinsic research performance of institutions.

In order to generalize the use of the h-index, it is important to understand the effect of the population size. In addition, while the h-index is a useful indicator for comparing

individuals' productivity in a given field, it should be noted that strong variations exist between disciplines and even between research topics within a given field. This was noted recently by MG Banks [7] who extended the interpretation of the h-index to help students judge the popularity of a given research topic. Clearly, as stated above, there is a danger of misinterpreting the h-index as a high h does not reflect only quality but also population size. Top scientists in Biology are more cited than top scientists in the Physics community, who are themselves more cited than top scientists in the Engineering community [8]. It is clear that these differences do not only reflect the quality of the work of individuals in questions but also reflect practices of a given field (or the population size of a scientific community).

With these warnings in mind, we conduct field *experiments* using the ISI Web of Knowledge database [1]. This database imposed an upper limit on the sample population of 10^5 papers.

Results

Fig. 1 shows the h-index in terms of number of papers for three renowned journals, Science (general interest), Acta Materialia (Materials Science journal, which was previously known as Acta Metallurgica), and the Journal of the Mechanics and Physics of Solids (Mechanics). Each point in this log-log diagram corresponds to a given country or a group of countries. For instance the USA has published a total number of 44700 papers in Science, and the h-index (at the date of November 26, 2006) of this set of paper is 719.

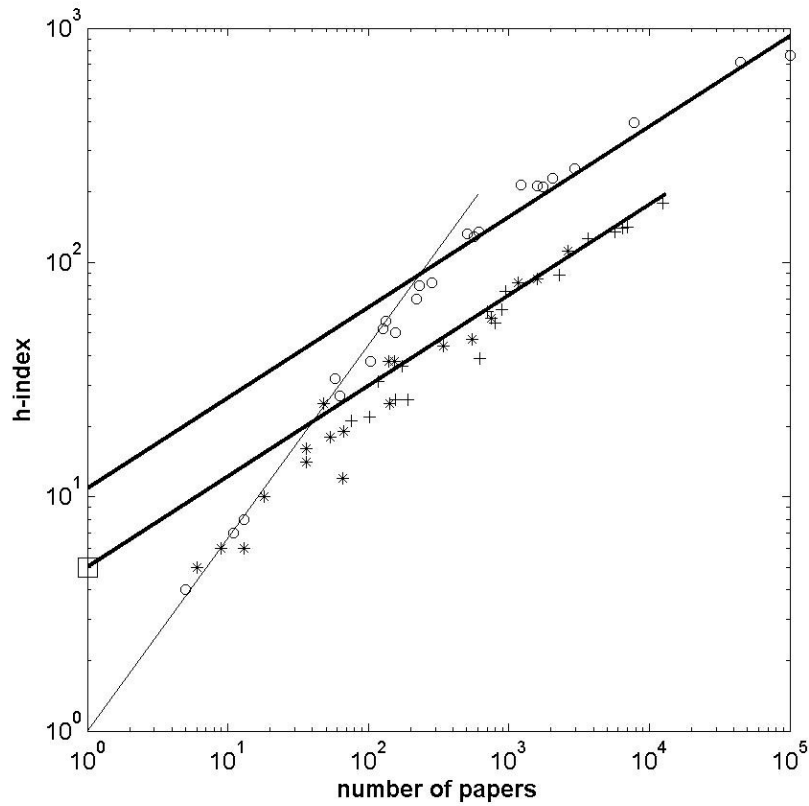


Fig. 1: master curves of scientific journals: Science (dots), Acta Materialia (crosses), the Journal of the Mechanics and Physics of Solids (asterisks). These master curves show the evolution of the h -index in terms of number of papers. Each point in this log-log diagram corresponds to a given country. The thin and bold straight lines are a fit to the data and distinguish two different growth regimes. The square symbol representing the intersection of the lowest bold line with the vertical axis characterizes h_m for Acta Materialia (or for the Journal of the Mechanics and Physics of Solids).

If, for a given journal (say Science) one considers various countries, it is remarkable to observe that all points form a line. This line, which will be shown to be similar in shape for all journals and for all set of papers published by research institutions, is called

henceforth the master curve (m-curve) of the journal (or institution) considered. Indeed, it can be shown that to a given set of papers S , one can associate in a unique way a theoretical reference m-curve defined by statistical treatment of the h-index of subsets randomly sorted out of S (mathematical developments are given in Appendix and in [9]). The m-curve of the Journal of the Mechanics and Physics of Solids is shown in Fig.2. Here, S is the set of the 2638 papers published in this journal until November 26, 2006. It appears that the asterisks in Fig.1 referring to papers published by particular countries are distributed along the reference m-curve of this journal.

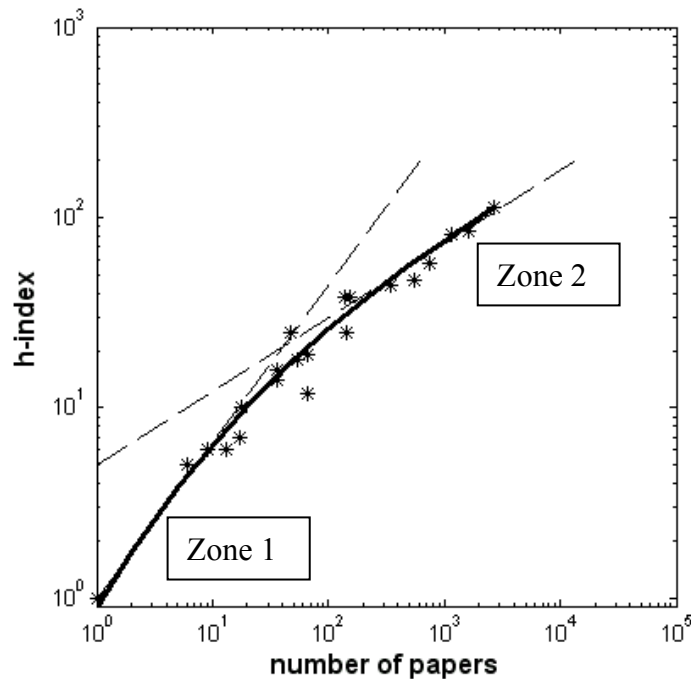


Fig. 2: theoretical master curve (m-curve) of the Journal of the Mechanics and Physics of Solids.

The theoretical m-curve is made of the assembly of discrete points (see Appendix), but is represented here, for clarity, as a continuous bold line. The dashed lines represent two different growth rates (zones 1 and 2). Asterisks are individual countries' scientific production for this specific journal and are distributed around the m-curve.

Each m-curve can be decomposed in two zones in the log-log diagram of Figs 1 and 2. In Fig.1, the thin straight line (zone 1 for small number of papers) and bold straight line (zone 2 for large number of papers) reflect two different growth rates. Zones 1 and 2 are also shown in Fig.2. For a large number of papers, the growth rate of the h-index is governed by a power law with exponent β (slope of the bold straight lines in Fig.1):

$$h = h_m N^\beta \quad (1)$$

where N is the number of papers. The value of β is about 0.4. A key observation we have made, is that this growth rate applies for all the institutions and journals we have considered (including those that are not shown in this paper) and has therefore a general character, at least for the range tested here, i.e. $N < 10^5$. Note that the level of zone 2 in Fig.1 is characterized by the intersection $\log(h_m)$ with the vertical axis (square symbol).

All the results obtained for Science, Acta Materialia and the Journal of the Mechanics and Physics of Solids, merge within zone 1, Fig.1. While the last two journals almost coincide within zone 2 (similar h_m), Science is characterized by a higher level of zone 2 (higher h_m). The value of h_m can be viewed as a measure of the impact or visibility of the papers published by a given journal. A point above the theoretical reference m-curve of a given journal is an indication of a better quality of the scientific production published in the journal by the country considered. Since for each journal the dispersion in figure 1 is not too large it is seen that all three journals considered here proceed to a fair selection of

papers based on quality, independently from the countries from which the papers originated.

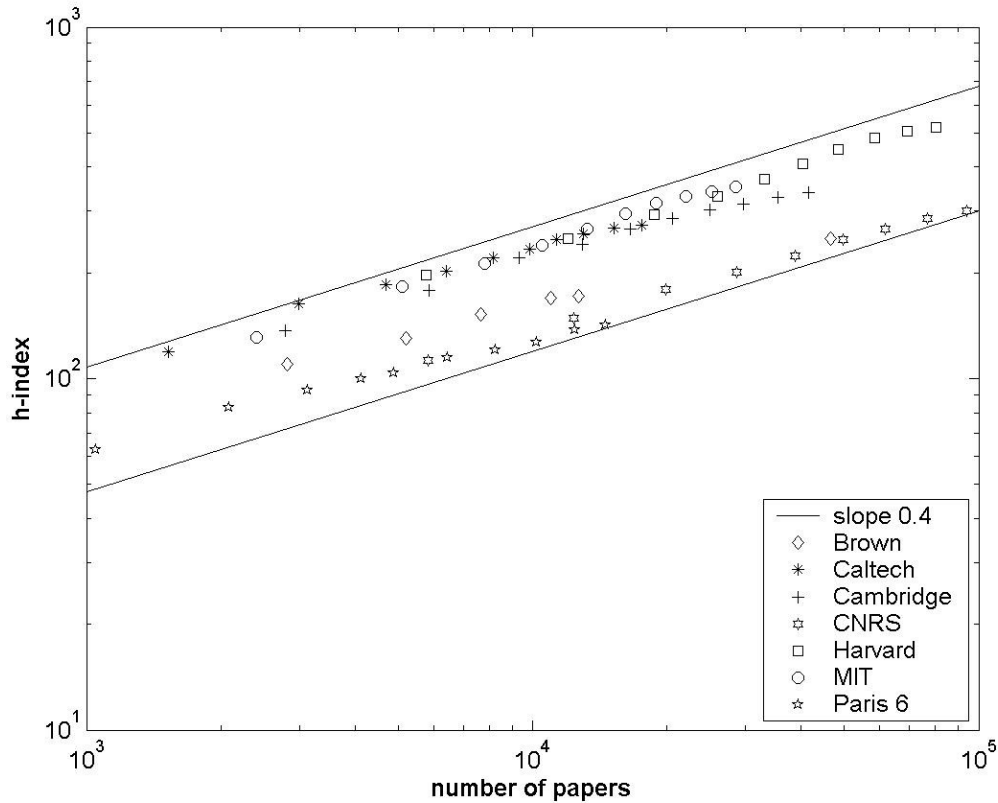


Fig. 3: h-index as function of population size for various international research institutions; note that for the large numbers of papers considered the h-index lies within zone 2. These results exemplify the universality of the power law exponent (around 0.4). The vertical shift between institutions is an indication of their respective visibility. The sample sizes were built by assembling papers published in even years starting in 1980 up to 1998 (more recent papers were not included as they are generally not sufficiently cited to impact the h-index). The two continuous lines with slope 0.4 are added on the figure to show the general growth rate but are in no means universal upper or lower bounds. Indeed, most universities would fall under the lower straight line.

Another illustration of the dependence of h on number of papers is given in Fig. 3 in which we compare academic institutions. We consider American and European institutions of excellent standing. These universities end up consistently on top of American and international rankings [10] (note that international rankings are strongly based on research productivity). If we single out Caltech and Harvard, one would experience difficulty at differentiating the quality of these institutions, yet Harvard's h -index for any given year is larger than Caltech's. Clearly, population size is here a strong factor. To understand this effect, we have superimposed the h -index for increasing population sizes for Harvard and Caltech on Fig.3. The database was collected by measuring (at the date of October 21, 2006) the h -index for all papers issued in 1980, then for the combined years 1980 and 1982, up to including all even years from 1980 to 1998. This procedure permitted the assembly of large set of papers. Amazingly, the h -indices as function of number of papers for these two universities align almost perfectly, which we dare to interpret as indication of equivalent research impact. On the same figure, we have added the data (obtained through the same procedure) for other international institutions. Cambridge research output seems of comparable visibility, whereas the impact index h_m emanating from France top university in the Shanghai University ranking [10] is slightly lower. It should be added that French national research institution CNRS has an output comparable to Paris 6. As a simple mathematical consequence of the power growth law, combining the paper populations of Paris 6 and

CNRS would increase the length of the straight line in Fig.1 but would not shift it vertically.

From these field observations, major comments can be drawn. First, the master curve associated to a given population of papers can be viewed as a signature of the visibility of these papers. For a large population, the corresponding h-index can be factorized as the product of an impact-index h_m with a size dependent factor $N^{0.4}$, where N is the population size. This decomposition illustrates the major feature of the h-index of being a metric accounting for both quality and quantity. The power law exponent of the order of 0.4 which governs the ultimate growth rate of the master curves is amazingly consistent between different institutions, and disciplines (as represented by different journals in Fig. 1). This exponent must convey a rate of diffusion of information and is in itself an interesting topic to analyse. Second, an assessment of the intrinsic visibility of an institution can be inferred from the vertical shifts of the curves in Figs. 1 and 2, which are characterized by h_m . An important feature of the index h_m associated to a given set of papers S is that it can be fairly well estimated by using any representative subset of S containing a sufficiently large number of papers. This property may be quite useful, since it is sometimes difficult to collect the totality of the papers published by some institutions, due to a lack of systematic classification or labelling.

In order to further establish the robustness of the h_m indicator for assessing research impact, we analyse in Fig. 4 its evolution with time for the three previous journals (Acta Materialia, Journal of the Mechanics and Physics of Solids, and Science). Here, h_m is

calculated for different years, with for example h_m (1980) corresponding to the set of papers published by a given journal during the year 1980. h_m is a decreasing function of time for recent years, which reflects that recent papers have not had yet enough time to accumulate citations. However h_m quickly reaches a plateau for Acta Materialia and for the Journal of the Mechanics and Physics of Solids, indication that the visibility of these journals is stable along the years considered. The evolution of h_m with time for Science has a different trend. Indeed, h_m appears to be increasing over the years 1976-1994, which we may interpret as a sign of reinforcement of the journal impact with time perhaps partly due to the increasing influence of expanding disciplines. Clearly, similar observations could be made for universities whose research is emerging, and reversely for institutions or departments that are losing ground (in that case h_m would not plateau but would decrease steadily over the years).

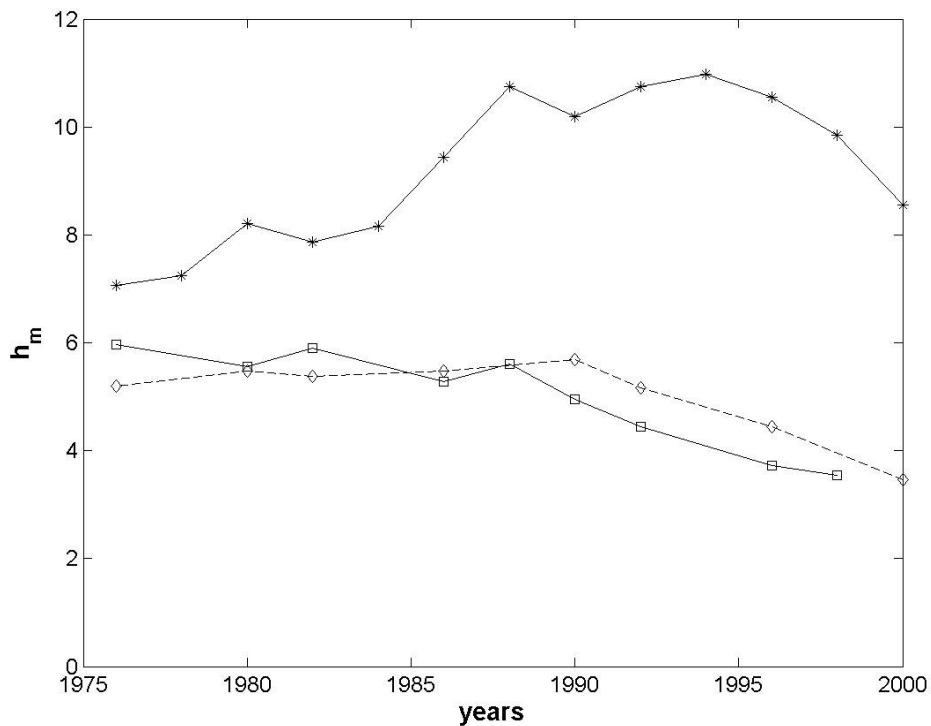


Fig. 4: Evolution of h_m with publication date for Science (stars), Acta Materialia (squares) and the Journal of the Mechanics and Physics of Solids (diamonds).

The index h_m may be put in perspective with the Highly-Cited (HiCi) criterion that enters in many rankings. A new classification that has gained high visibility is the Shanghai's university 2006 international ranking [10]. One of the criteria entering in [10] is the number of Highly-Cited researchers in a given institution, which may advantage large universities. Our computation of h_m for a group of scientific domains (see Table 1) and for selected universities entering in the top 100 positions in [10] shows how it correlates with HiCi. For instance, Harvard University remains in top position. However, small yet excellent institutions such as Caltech and Princeton appear reinforced. Thus, h_m appears to correct the size effect and its use is more general than Hi-Ci, as it can be extended to compare scientific journals or institutions and laboratories of small sizes. As an additional test we consider the grouping of two institutions (a solution sometimes proposed to increase the visibility and ranking by increasing the size). The case study we consider here is the merging of Berkeley (USA) and Cambridge (England), two universities with almost identical h_m . The union of these institutions yields for the research domains considered in Table 1 a total of 10480 papers. Note that this number is slightly lower than the direct sum of the papers produced by both institutions, which is an indication of the volume of collaborative work. The h-index of the union is 152, a number larger than the h-index of each institution (115 and 116). However, h_m is left unchanged at 3.75. A mathematical formalism for merging rules is proposed in a follow up work [9].

This example clearly illustrates the interest of using a size-corrected index such as h_m to evaluate the intrinsic performance of a research institution or journal.

Conclusions

The signature of the visibility of a given population of papers can be characterized by a single curve (called master-curve or m-curve). This curve shows the growth of the h-index in terms of the number of papers. For a large population (more than some hundreds) the level of the master curve is defined by a single indicator h_m characterizing the impact of the papers considered. This indicator h_m has the important feature of being size-corrected. This was exemplified by the case study of the merging of two universities of same h_m with the effect of leaving the combined h_m unchanged. The h-index of a large population of papers can be decomposed into the product of the impact index h_m with a factor depending on the population size.

Due to its simplicity, in complement to existing indicators, we advocate using the impact index h_m for comparing journals within a given field and general research institutions (of different sizes). It could be used as well to compare relatively small-size research groups such as departments, laboratories and institutes.

The mathematical foundation of the results presented in this paper is developed in a follow up paper [9].

Institution name	HiCi score	<i>N</i> : number of papers	h-index	$h_m = h/N^{0.4}$ (Table rank)
Harvard University	100 (1)	4280	140	4.94 (1)
Stanford Univ	88.4 (2)	3381	111	4.30 (5)
Univ California – Berkeley	70.5 (3)	5294	116	3.76 (10)
Massachusetts Inst Tech	66.6 (4)	4541	127	4.37 (4)
Princeton Univ	59.6 (5)	3454	117	4.50 (3)
California Inst Tech	59.1 (6)	2751	109	4.56 (2)
Columbia Univ	56.0 (7)	1625	78	4.05 (7)
Univ Cambridge	53.8 (8)	5244	115	3.74 (11)
Univ Chicago	49.9 (9)	1626	82	4.26 (6)
Univ Oxford	48.0 (10)	3801	99	3.66 (14)
Duke Univ	48.0 (11)	1711	76	3.87 (9)
Johns Hopkins Univ	40.7 (12)	2411	89	3.95 (8)
Michigan State Univ	37.7 (13)	1340	66	3.70 (12)
Univ Utrecht	28.8 (14)	1402	64	3.53 (16)
Osaka University	25.5 (15)	563	37	2.94 (19)
Paris 6	23.1 (16)	2465	67	2.95 (18)
Case Western Reserve Univ	21.8 (17)	1038	59	3.67 (13)
Univ. Birmingham	21.8 (18)	1426	60	3.28 (17)
Univ Munich	15.4 (19)	3296	91	3.56 (15)

Univ Roma – La Sapienza	10.9 (20)	1789	56	2.80 (20)
-------------------------	-----------	------	----	-----------

Table 1: Selected universities from the 2006 Shanghai's university ranking [10] here ranked by their Hi-Ci score (related to the number of highly cited researchers within the institution). The chosen universities were in the first ten ranks in [10] plus ten additional institutions separated by about ten positions up to rank 100. The number of papers was obtained through a Web Of Science search for the 1994-2003 time period and for the following grouping of disciplines: Physics, Mathematics, Chemistry, Materials, Engineering, Mechanics. Note that all papers may not have been selected for each universities because of a lack of systematic classification; however, each obtained set is larger than some hundreds which yields a safe determination of h_m . Note also that the results of fig.2 were obtained by considering all disciplines, with the effect of producing a higher index h_m as a consequence of the higher level of citations for papers in medicine and biology.

References and Notes:

1. ISI Web Of Knowledge, <http://portal.isiknowledge.com>
2. J.E. Hirsch, *Proc. Nat. Acad. Science*, **102**, 16569 (2005).
3. P. Ball, *Nature*, **436**, 900 (2005).
4. G. Saad, *Scientometrics*, **69** (1), 117 (2006).
5. W. Glanzel, *Scientometrics*, **67** (2), 315 (2006).
6. T. Braun, W. Glanzel, A. Schubert, *Scientometrics*, **69** (1), 169 (2006).
7. M.G. Banks, *Scientometrics*, **69** (1), 161 (2006).
8. http://en.wikipedia.org/wiki/Hirsch_number

9. J.F. Molinari, A. Molinari, “Mathematical foundation of a new criterion for ranking scientific institutions based on the h-index”, to be submitted at *Scientometrics* (2007).
10. Shanghai’s 2006 International university ranking:
<http://ed.sjtu.edu.cn/ranking.htm>

Appendix: definition of the m-curve

In this appendix, the relationship between the h-index and number of papers is analysed.

Consider a set S of papers. The h-index of S is the number of papers h of S which are cited at least h times. The h-index of S is denoted by $h(S)$.

A.1 Citation law

Let us denote by $N_T = \text{card}(S)$ the total number of papers in S , and by $i \rightarrow p(i)$ ($i=1,2,\dots$

N_T) a permutation of the papers of S ; $N_c(i)$ is the number of citations of the paper $p(i)$.

The permutation $p(\cdot)$ is chosen such that $N_c(i)$ is decreasing with i . In other words the papers are sorted in descending number of citations, $N_c(1) \geq N_c(2) \geq \dots \geq N_c(N_T)$. Note that some of the $N_c(i)$ may have identical values.

The set of papers S is characterized in the present analysis by the function $N_c(\cdot)$:

$i \rightarrow N_c(i)$ defined on $(1,2,\dots, N_T)$, which is called henceforth the *citation-law* of S .

The number n_d of distinct values of $N_c(i)$ is smaller or equal to $N_c(1)$. Let us sort the

distinct values of $N_c(i)$ in descending order $(N_c(G_1), N_c(G_2), \dots, N_c(G_{n_d}))$, with $G_1 = 1$,

$G_1 < G_2 < \dots < G_{n_d}$ and $N_c(G_i) \neq N_c(G_j)$ for $i \neq j$. For each m ($1 \leq m \leq n_d$) F_m is defined as

the largest integer such that $N_c(F_m) = N_c(G_m)$; F_m is also the number of papers having a number of citations larger or equal to $N_c(G_m)$.

A.2 Dependence of the h-index in terms of the number of papers: the master curve

Consider two integers n and c such that $n \leq N_T$ and $c \leq n$. Denote by S_n a combination of the papers of S , n at a time. The total number of these combinations is $\binom{N_T}{n}$. The

notation $\binom{p}{q} = \frac{p!}{q!(p-q)!}$ is used to represent the number of combinations of p elements, q at a time.

Denoting by $\tilde{N}(n, c)$ the number of combinations S_n which have an h-index equal to c ($h(S_n) = c$), we have:

$$\binom{N_T}{n} = \sum_{c=0}^n \tilde{N}(n, c) \quad (\text{A1})$$

The probability $P_r(n, c)$ to have a combination S_n such that $h(S_n) = c$ is:

$$P_r(n, c) = \frac{\tilde{N}(n, c)}{\binom{N_T}{n}} \quad (\text{A2})$$

From (A1), it follows that $\sum_{c=0}^n P_r(n, c) = 1$.

Clearly we have $\tilde{N}(n, c) = P_r(n, c) = 0$ for $c > N_c(1)$, or for $c > n$.

For $c \leq N_c(1)$ there is an unique integer i such that $c = N_c(G_i)$ or

$N_c(G_{i-1}) > c > N_c(G_i)$. Considering these two possibilities, it can be shown that, for any

integer n such that $c \leq n \leq N_T$, $\tilde{N}(n, c)$ is given by:

$$\tilde{N}(n, c) = \sum_{k=0}^{a_i} \sum_{j=k}^{b_i} \binom{F_{i-1}}{c+k-j} \binom{b_i}{j} \binom{N_T - F_i}{n-c-k} \quad \text{for } c = N_c(G_i), \quad (\text{A3})$$

with:

$$a_i = \inf(n - c, b_i) \quad b_i = F_i - F_{i-1}, \quad (\text{A4})$$

and by

$$\tilde{N}(n, c) = \binom{F_{i-1}}{c} \binom{N_T - F_{i-1}}{n-c} \quad \text{for } N_c(G_{i-1}) > c > N_c(G_i) \quad (\text{A5})$$

Note that (A3) can also be written as:

$$\tilde{N}(n, c) = \binom{F_i}{c} \binom{N_T - F_i}{n-c} + \sum_{k=1}^{a_i} \sum_{j=k}^{b_i} \binom{F_{i-1}}{c+k-j} \binom{b_i}{j} \binom{N_T - F_i}{n-c-k} \quad \text{if } c = N_c(G_i) \quad (\text{A6})$$

With Eqs. A3-A5, one can now determine all relevant statistical information including

the mean h -index for the combinations S_n :

$$\bar{h}(n) = \sum_{h=0}^n h P_r(n, h) \quad (\text{A7})$$

and the variance and standard deviation:

$$\text{var}(h, n) = \sum_{h=0}^n (h - \bar{h})^2 P_r(n, h) \quad \sigma(n) = \sqrt{\text{var}(h, n)} \quad (\text{A8})$$

The master curve (or m-curve) of the set of papers S is by definition the curve $n \rightarrow \bar{h}(n)$

$(1 \leq n \leq N_T)$.