Made available by Hasselt University Library in https://documentserver.uhasselt.be

Thoughts on Uncitedness: Nobel Laureates and Fields Medalists as Case Studies

Peer-reviewed author version

EGGHE, Leo; Guns, Raf & ROUSSEAU, Ronald (2011) Thoughts on Uncitedness: Nobel Laureates and Fields Medalists as Case Studies. In: JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 62(8), p. 1637-1644.

DOI: 10.1002/asi.21557
Handle: http://hdl.handle.net/1942/12894

Thoughts on uncitedness: Nobel laureates and Fields medalists as case studies

Leo Egghe[1,2], Raf Guns[2] and Ronald Rousseau [1,2,3,4]

[1] Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan,
B-3590 Diepenbeek, Belgium
E-mail: leo.egghe@uhasselt.be

[2] Universiteit Antwerpen (UA), IBW, Stadscampus, Venusstraat 35,
B-2000 Antwerpen, Belgium
E-mail: raf.guns@ua.ac.be

[3] KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101,
B-8400 Oostende, Belgium
E-mail: ronald.rousseau@khbo.be

[4] K.U.Leuven, Dept. Mathematics,
Celestijnenlaan 200B, B-3000 Leuven (Heverlee), Belgium

ABSTRACT

Contrary to what one might expect, Nobel laureates and Fields medalists have a rather large fraction (10% or more) of uncited publications. This is the case for (in total) 75 examined researchers from the fields of *mathematics* (Fields medalists), *physics*, *chemistry*, and *physiology or medicine* (Nobel laureates). We study several indicators for these researchers, including the h-index, total number of publications, average number of citations per publication, the number (and fraction) of uncited publications, and their interrelations. The most remarkable result is a positive correlation between the h-index and the number of uncited articles. We also present a Lotkaian model, which partially explains the empirically found regularities.

Keywords and phrases: uncitedness, Nobel Prize, Fields Medal

Introduction

It took us by surprise to notice that a winner of the Fields medal (the highest award in mathematics, named after John Charles Fields and awarded by the Fields Institute) had a non negligible percentage of uncited publications. Checking some other awardees revealed that this was not an exception. For some of them this percentage turned out to be more than 10%, excluding editorials or book reviews. We were surprised because we expected that top scientists, and especially mathematicians, only write publications dealing with difficult and important problems, which, once solved, lead to high quality and hence highly cited publications. Assuming that a publication by such a visible author is not his best work, even then one expects that, because of his status, this work would not be ignored. In other words, one expects the Matthew Effect at work. This observation led to an investigation of uncitedness among outstanding researchers. More particularly, we looked at Fields medalists and Nobel Prize winners.

We investigated the field of *mathematics* (Fields medal winners: in principle four winners every four years) and Nobel Prize winners in the fields of *physics*, *chemistry* and *physiology or medicine* (one, two or three winners a year). It became immediately clear that even among these eminent scientists having many uncited publications was quite common. In most cases, we found more than 10% uncited publications and this over all studied fields. Of course, an uncited publication may, in principle, gain citations at some time in the future. This is related to the phenomenon of *Delayed Recognition* (Garfield, 1980; Glänzel, Schlemmer and Thijs, 2003). Publications that remain uncited for a prolonged period of time and subsequently receive several citations are known under the name of *Sleeping Beauties* (van Raan, 2004; Burrell, 2005). In other words, such an – as yet – uncited publication is not necessarily a 'never' cited publication, see the methodological part in the next section.

This phenomenon (the uncited publications of top researchers) has apparently not yet been addressed in the literature, although Glänzel et al. (2006) note in passing: "The fact that a document is less frequently cited or even (still) uncited several years after publication provides information about its reception by colleagues but does not reveal anything about its quality or the standing of its author(s) in the community. Uncited papers by Nobel Prize winners may just serve as an example."

Data were collected during the period October-November 2010 using Thomson Reuters' Web of Science (WoS). While collecting the total number of publications and the number of uncited publications, we also collected the readily available average number of citations per publication and the author's h-index (Hirsch, 2005).

Next we studied the scatter plots resulting from the relations between any two of the above mentioned indicators. We especially looked for increasing or decreasing

relationships. This revealed (details are given in the next section) an increasing relationship between total number of publications and the h-index, the h-index and the average number of citations per publication, the number of uncited publications and the total number of publications and between the number of uncited publications and the h-index. Decreasing relationships were found between the number of uncited publications and the average number of citations per publication, and between the fraction of uncited publications and the average number of citations per publication.

In the third section, we examine how these indicators are related in a continuous Lotkaian system. The Lotka system can be considered a first approximation of the reality that publications are cited in a much skewed way. In this setting, one can prove most of the decreasing and increasing relationships that were found empirically (sometimes needing an extra condition). This yields a partial explanation (because we assume Lotkaian systems) of the above findings. Although the Lotkaian system has some drawbacks (it lacks the 0 as frequency and is only an approximation of reality), we believe that it is a good approximation that still allows for partial explanation of the relations found: the mathematics of more intricate models quickly grows too complicated.

The paper ends with conclusions, open problems and other suggestions for further research.

Methodology

We obtained the list of recent Fields medalists from the website of the International Mathematical Union (http://www.mathunion.org/general/prizes/fields/prizewinners/) and recent Nobel laureates in *physics*, *chemistry* and *physiology or medicine* from the website of the Nobel Prize committee (http://nobelprize.org/). We restricted the number of laureates to a number between 15 and 26 per discipline in order to find clear relations between any of the two indicators mentioned above and to obtain manageable clouds of points. Bringing all fields together already yields clouds of 75 points. In practice we included 18 Fields medalists (mathematics) between 1990 and 2006; 16 Nobel laureates in physics between 2004 and 2010; 15 Nobel laureates in chemistry between 2004 and 2010; and 26 Nobel laureates in physiology or medicine between 1999 and 2010. Scientists with very common names were not included as it was too difficult to collect correct data. We consider it acceptable to delete a few names and think that this does not bias the data set used by us. Citation data collection took place during the period October-November 2010 from Thomson Reuters' Web of Science (WoS, including Proceedings).

Using the advanced search facility, queries were performed as follows. First we did a search on the author's name followed by the first initial and an asterisk. This often

revealed a second initial so that a specific query could be made. If such a specific query was possible, we searched for the author with one or two initials (no asterisk anymore), otherwise we just continued with the original result. The result was then limited to 'possible' subfield categories. This list was then "analyzed" for institutes. The result showed the institute or institutes where the scientist worked during his/her career and, more importantly, revealed homonyms working in the same field. These were deleted.

The following data were collected:

T_1: the total number of publications

n_1: the total number of uncited publications

μ: the average number of citations per publication

h: the h-index

The average number of citations per publication was obtained from WoS' citation report. The number n_1 (the total number of uncited publications) may refer to many recent publications. Further on we will, however, only consider uncited publications published before the year 2006, since it is very unlikely that they will gain any citations later on. These publications will be called "never-cited" publications.

Of course, one can never be absolutely certain that these will never be cited, but the time period used (2005 or older) guarantees that most of these publications will indeed be never-cited. To verify this claim, we examined all publications (n = 332) published in 1990 in the five journals that were ranked highest in the JCR category 'Biology' (ranked according to the impact factor). It was found that 70 publications had not yet been cited after 5 years; of these, only 4 (5.7%) gained citations in later years. Moreover, their citation numbers are rather low: 1, 1, 2, and 8 respectively (in January 2011). This small case study illustrates how extraordinary it is for publications to gain a first citation after more than five years. Indeed, as shown theoretically by Burrell (2002), the longer a publication remains uncited, the less likely it is to ever gain a citation. We therefore conclude that virtually all "never-cited" publications will indeed never be cited.

The never-cited publications constitute the real objective of our study. Therefore we also collected the following data, by appropriately limiting publication years:

T_2: the total number of publications published strictly before 2006

n_2: the total number of never-cited publications, i.e. those publications included in T_2 which were uncited at the moment of data collection.

We will not only study the never-cited indicators n_2 and n_2/T_2 (the fraction of never-cited publications) in relation to the other indicators, but also the relations between h, T_1, T_2

and μ. Note that we do not use versions of h or μ that are restricted to publications before 2006. We found out that the results are the same qualitatively and even quantitatively as there is almost no difference between the two "μ"-versions and the h-index just stays unchanged in almost all cases (because recent publications usually do not contribute to the h-index). Results are described in the next section.

Empirical results

Fig. 1 depicts the relation between h and T_2 (Fig. 1). The corresponding scatter plot relating h and T_1 is basically identical and hence not shown. For all fields these plots show an increasing relationship and a roughly concave shape. They correspond to the results shown in (Liu, Rao & Rousseau, 2009) for horticulture journals. In the next section we will show how this shape can be explained assuming a Lotkaian distribution.

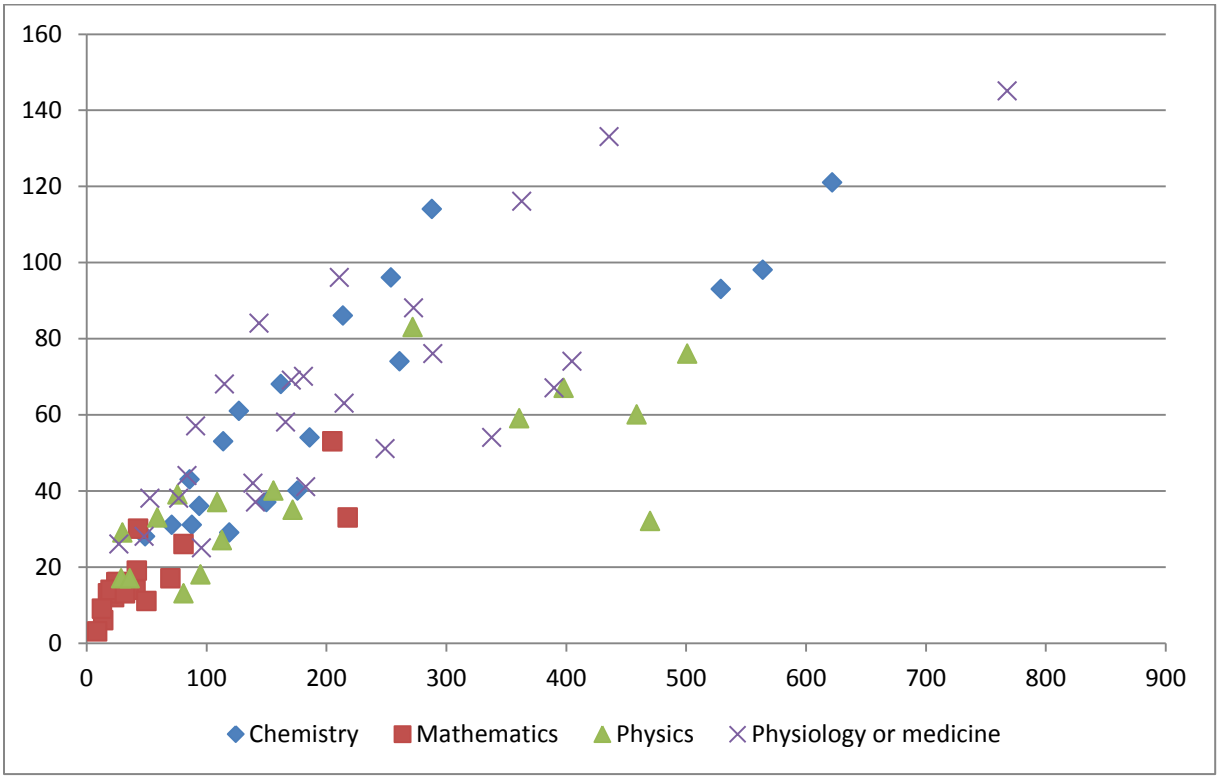

Fig. 1 Empirical relation between T_2 (horizontal axis) and h (vertical axis)

The relation between h and μ is depicted in Fig. 2. Although Fig. 2 is more scattered than the previous figure we can still see the (expected) increasing relation between h and μ. This will be partially explained in the next section.

Now we come to the main topic of this contribution, namely uncitedness, or better never-citedness. Fig. 3 shows box plots of the relative number of uncited publications per field. These illustrate that the median lies around 10 to 20% of papers that are uncited. We now turn to the relations between uncitedness and other indicators. Fig. 4 clearly shows the increasing relation between n_2 (total number of never-cited publications) and T_2 (total number of publications published strictly before 2006). This is an interesting result: the more publications an author has, the more never-cited publications (in general). This scientometric observation, which we will partially explain in the next section, provides a rationale for the fact that also highly visible scientists, such as Nobel laureates, can have many never-cited publications. If we assume that some percentage of all publications will remain uncited, an increasing relation between number of never-cited publications and total number of publications automatically follows. Of course, this is an explanation based solely on descriptive statistics. A complete explanation would have to take the content and potential impact of these never-cited publications into account; this is, however, beyond the scope of the present paper.

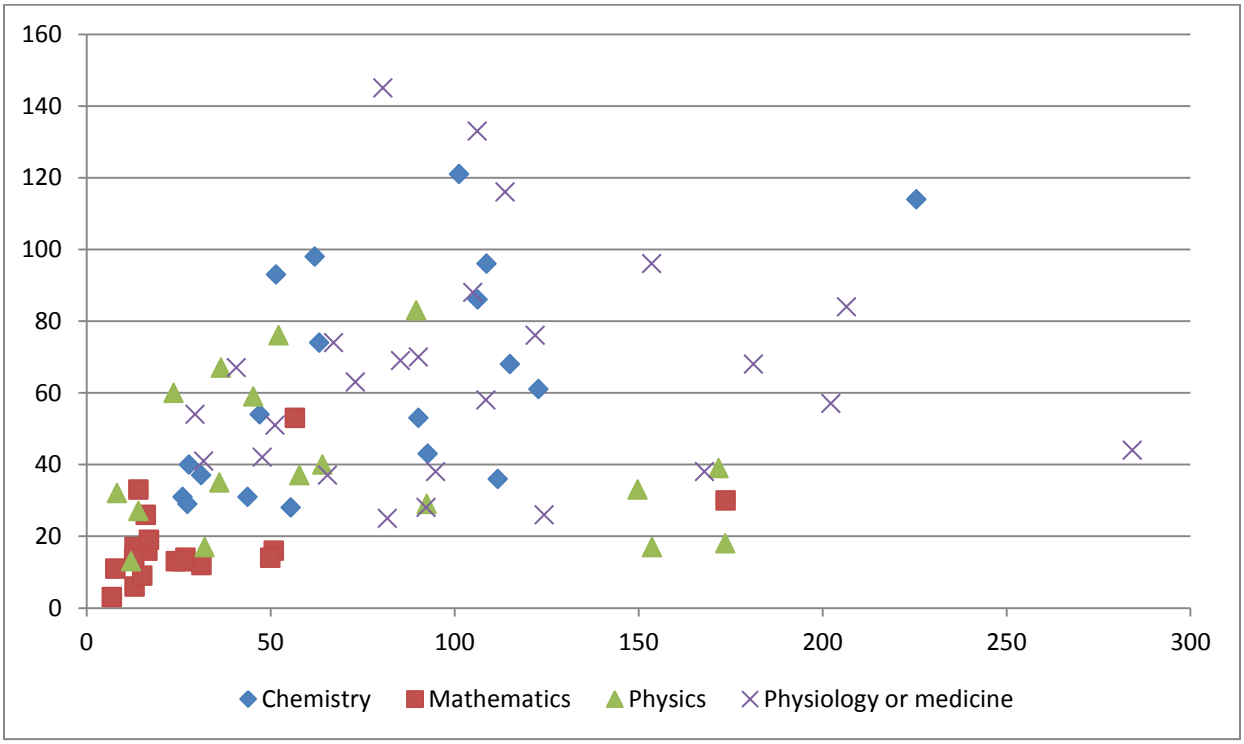

Fig. 2. Empirical relation between h (vertical axis) and μ (horizontal axis)

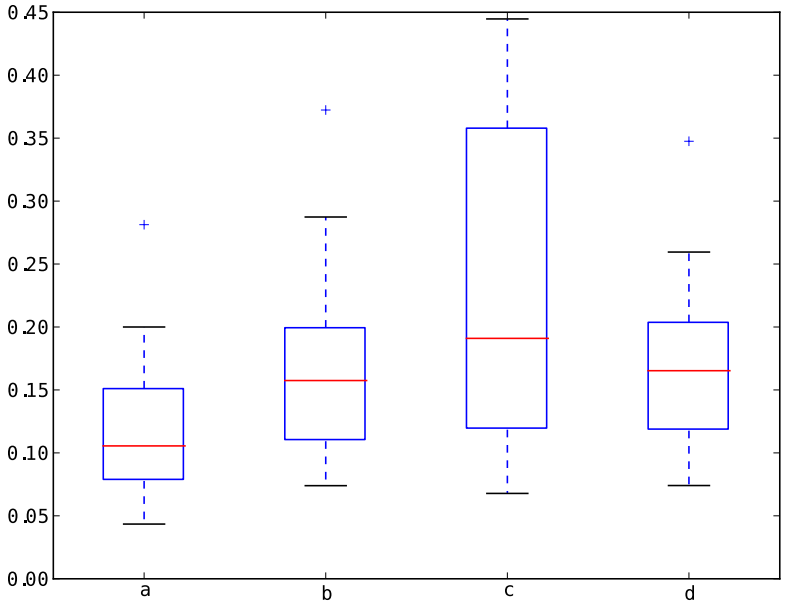

Fig. 3. Box plot of relative uncitedness (n_2/T_2) per field
(a: mathematics, b: chemistry, c: physics, d: physiology or medicine)

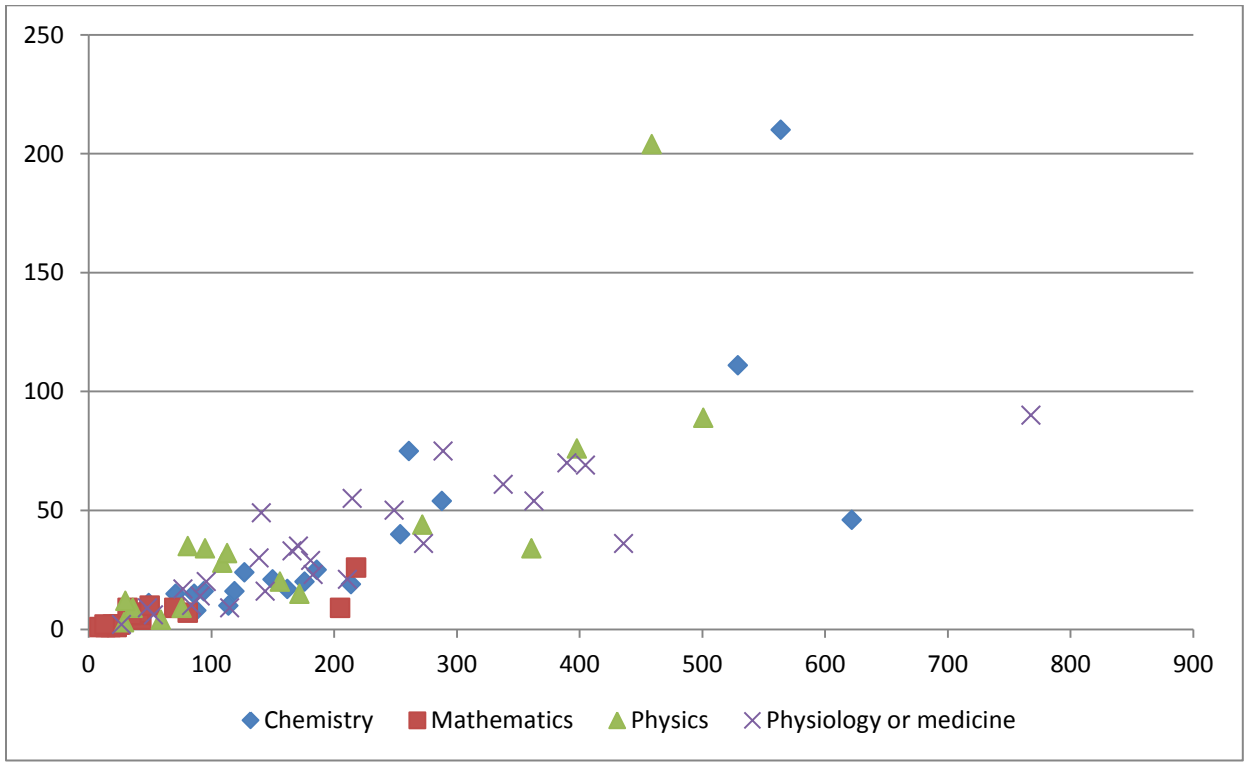

Fig. 4 Empirical relation between n_2 (vertical axis) and T_2 (horizontal axis)

Fig. 5 shows the relation between n_2 and μ. Although quite scattered, we may infer, at least visually, that this relation is decreasing. This visual observation is, however, not confirmed statistically, since the correlation coefficient is not significantly different from

zero. So, we must admit that a decreasing trend is weak at best. Yet, on logical ground, such a decreasing trend seems to be expected: the higher the (absolute) number of never-cited publications, the lower the average number of citations per publication (in general). In the next section, it will be shown that a decreasing relation is also expected in the Lotkaian model.

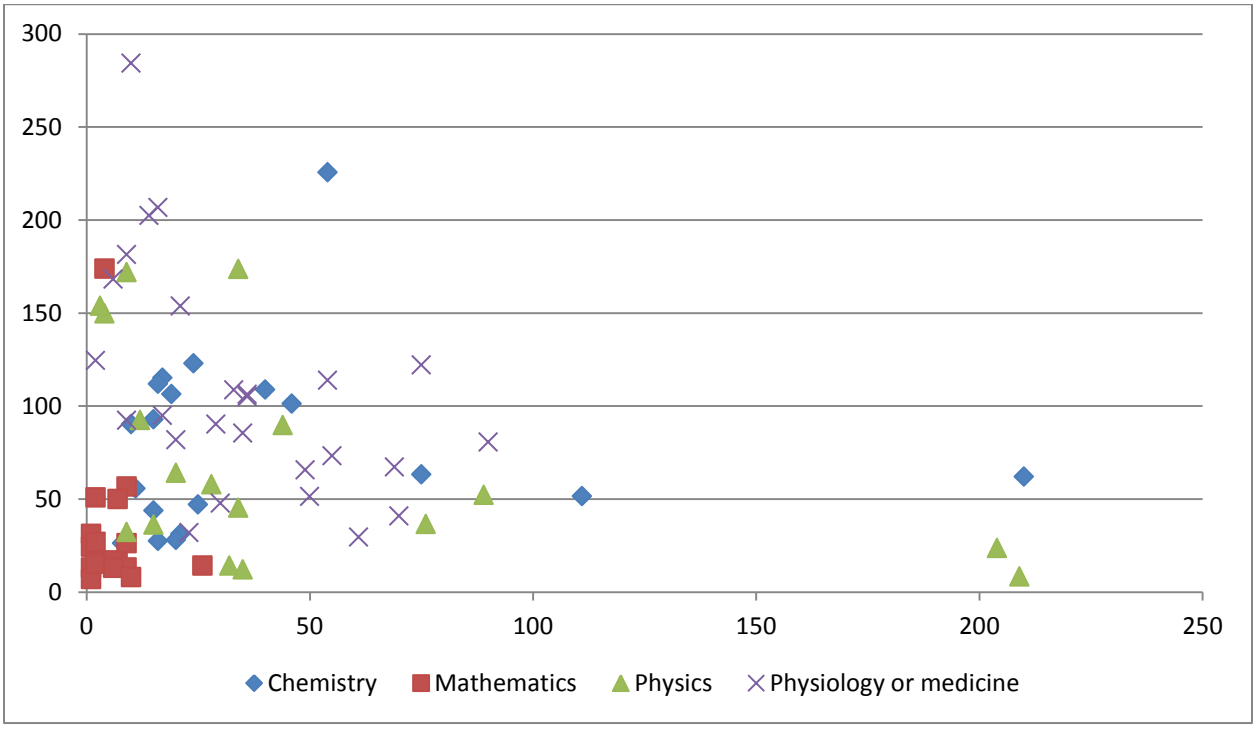

Fig. 5 An empirical relation between μ (vertical axis) and n_2 (horizontal axis).

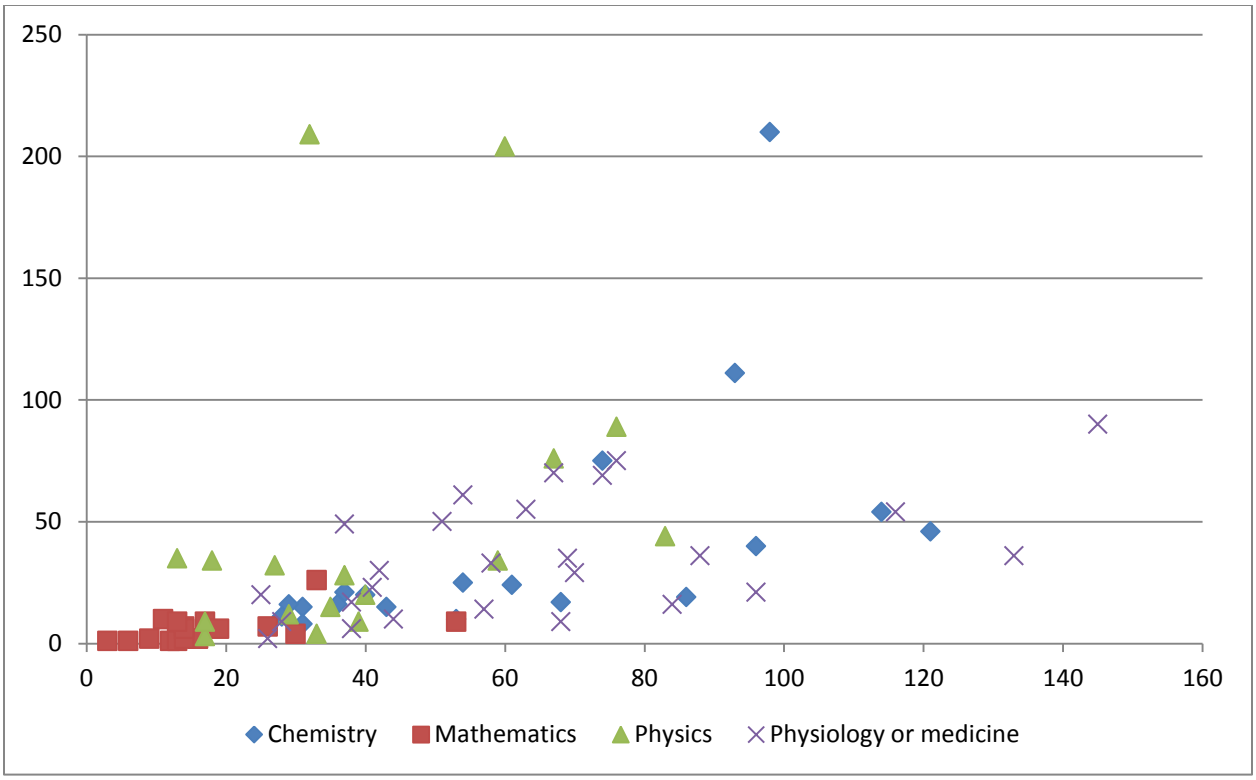

Fig. 6 An empirical relation between n_2 (vertical axis) and h (horizontal axis)

Fig. 6 depicts an increasing relation between n_2 and h. At first sight this might be surprising since an increasing h-index implies that publications are more cited and this should decrease n_2. However, this is not the case. Uncited publications never contribute to the h-index; hence there exists no direct relationship between the h-index and the number of uncited publications. The increasing relationship is due to the fact that both indicators are correlated with T: there exists an increasing relation between h and T_2 (Fig. 1) and an increasing relation between n_2 and T_2 (Fig. 4). Hence, based on these facts, the increasing relation shown in Fig. 6 is not surprising. Further explanations will be given in the sequel.

We further investigated the relation between the fractions of never-cited publications (fraction with respect to T_2): n_2/T_2 and T_2, n_2, μ and h. Only weak relations could be found. We only show (Fig. 7) the cloud of points depicting the relation between n_2/T_2 and μ. This shows a (weak) decreasing relation, which will also be proved in the next section. A referee suggested also showing the relation between n_2/T_2 and n_2. However, the resulting cloud of points is very scattered, and no conclusions could be drawn from it. We therefore do not include it.

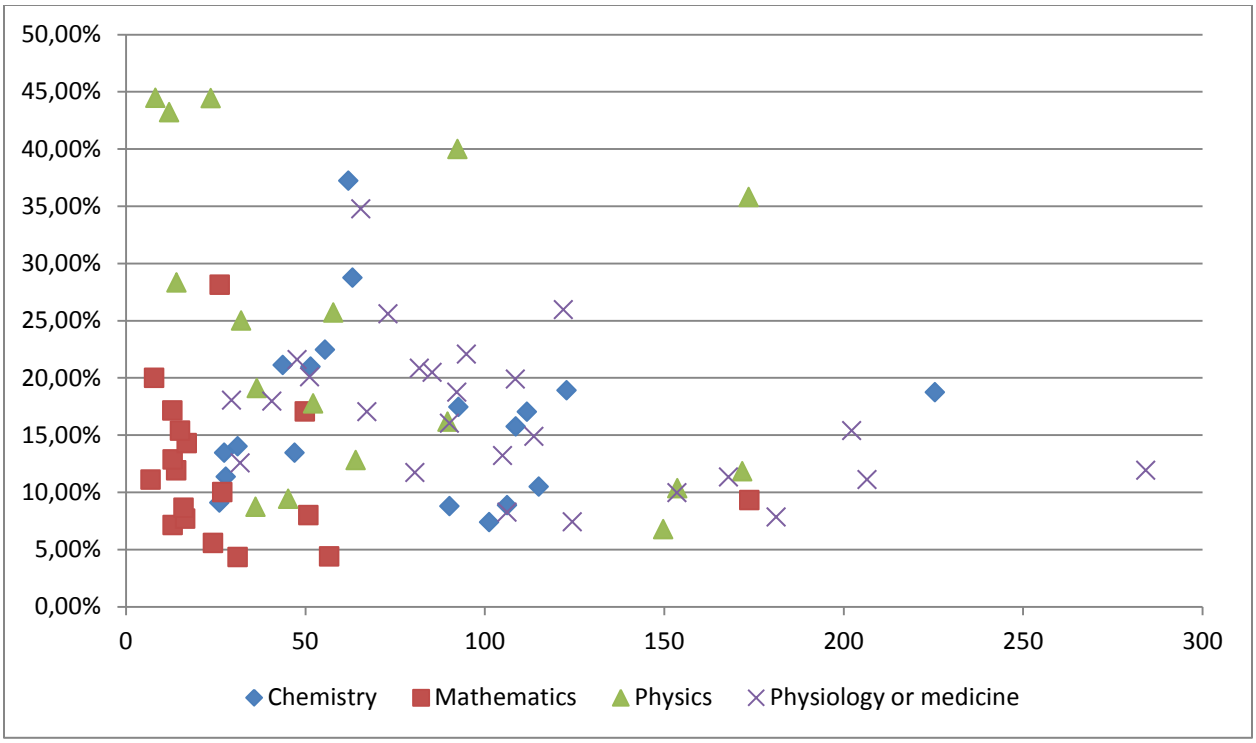

Fig. 7 Empirical relation between n_2/T_2 (vertical axis) and μ (horizontal axis)

Theory

In this section we will show that several empirical results of the previous section can also be found theoretically in a continuous Lotkaian framework. Here we do not make a distinction between “old” and “all” publications; a similar remark holds for citedness. We use the following notations:

T: total number of publications

μ: average number of citations per publication

h: h-index

n: total number of non-cited publications

We work in a Lotkaian framework, where the density of publications with citation density j is given by

$$f(j) = \frac{C}{j^{\alpha}} \tag{2}$$

with $C > 0$, $\alpha > 1$, $j \geq 1$. It is well-known that

$$T = \int_1^\infty f(j)\, dj = \frac{C}{\alpha-1} \quad (3)$$

(see e.g. Egghe (2005)), and if $\alpha > 2$,

$$A = \int_1^\infty j\, f(j)\, dj = \frac{C}{\alpha-2} \quad (4)$$

where *A* denotes the total number of citations received by these *T* publications. Hence

$$\mu = \frac{A}{T} = \frac{\alpha-1}{\alpha-2} \quad (5)$$

or, equivalently:

$$\alpha = \frac{2\mu-1}{\mu-1} \quad (6)$$

In (Egghe & Rousseau, 2006) we proved the following formula for the h-index:

$$h = T^{1/\alpha} \quad (7)$$

This formula corresponds to a concave function, which explains the shape we empirically found in Fig. 1.

Combining equations (6) and (7) yields:

$$h = T^{\frac{\mu-1}{2\mu-1}} \quad (8)$$

Formula (8) connects the three indicators *T*, μ and *h*. This leads to a partial explanation of Figs. 1 and 2, given in Proposition 1.

Proposition 1. If μ and *T* are strictly increasing then *h* is strictly increasing. The same conclusion holds when one of the two parameters is constant and the other one is strictly increasing.

Proof. This result readily follows from equation (8) and the fact that for $\mu > 1$ (see eq. (5)) $\frac{\mu-1}{2\mu-1}$ is a strictly increasing, positive function of μ.

Similarly we have Proposition 2.

Proposition 2. If μ and T are strictly decreasing then h is strictly decreasing. The same conclusion holds when one of the two parameters is constant and the other one is strictly decreasing.

This is proved in the same way as Proposition 1. Details are left to the reader.

We approximate the number of non-cited papers, denoted as n, with the expression:

$$n = \int_1^2 f(j)dj \tag{9}$$

A referee remarked that (9) in fact approximates the number of papers with one citation rather than zero citations. This is correct: since $j \geq 1$, Lotkaian systems do not include the sources that produce 0 items (here: the uncited or never-cited papers). The theoretical analysis here is thus concerned with lowly cited rather than uncited publications. The main reason for using a Lotkaian system instead of other distributions that do include the zero is simplicity. We admit that (9) is, at best, a crude approximation, but our approach has the advantage of yielding relatively simple formulae (see also the final section).

Alternatively, in the empirical part, we could have studied the total number of lowly cited publications, instead of the non-cited ones. By lowly-cited publications we mean those with at most one citation (or at most 2 citations). We are convinced that for lowly cited publications very similar graphs would have emerged.

Formula (9) boils down to calculating:

$$n = \int_0^1 \frac{C}{(j+1)^\alpha}\, dj \tag{10}$$

now for $j \geq 0$. It is easily seen that

$$n = \frac{C}{\alpha-1}\left(1 - \frac{1}{2^{\alpha-1}}\right) \tag{11}$$

But, by (3), we have:

$$n = T\left(1 - \frac{1}{2^{\alpha-1}}\right) \tag{12}$$

Using equation (6), equation (12) becomes:

$$n = T\left(1 - \frac{1}{\frac{\mu}{2^{\mu-1}}}\right) \tag{13}$$

which shows the connection between the three indicators T, μ and n. Finally,

$$\frac{n}{T} = 1 - \frac{1}{\frac{\mu}{2^{\mu-1}}} \tag{14}$$

which brings us to proposition 3.

Proposition 3. If μ is strictly increasing, then n/T is strictly decreasing.

Proof. This result follows from equation (14) since $\frac{\mu}{\mu-1}$ is positive and strictly decreasing for $\mu > 1$.

Proposition 3 explains the decreasing trend observed in Fig. 7. We formulate two other propositions.

Proposition 4. If μ is strictly increasing and T is strictly decreasing then n is strictly decreasing. If one of the two variables (μ or T) is constant, then the same conclusion holds.

Proposition 5. If μ is strictly decreasing and T is strictly increasing then n is strictly increasing. If one of the two variables (μ or T) is constant, then the same conclusion holds.

The proofs follow along the same lines as the proof of Proposition 3, using formula (13).

Proposition 5 partially explains Fig. 4. The next proposition partially explains Figs. 5 and 6.

Proposition 6. If μ is strictly decreasing and h is strictly increasing then n is strictly increasing. If one of the two variables (μ or h) is constant, then the same conclusion holds.

Proof. By equation (8) and since $\frac{\mu-1}{2\mu-1}$ is a strictly increasing, positive function of μ we see that T is strictly increasing. Combining this result with Proposition 5 yields that n is strictly increasing.

Clearly other relationships can be found, but as we do not need them in this article this is left to the reader.

Conclusions and open problems

This paper shows that the group of Fields medalists (mathematics) and Nobel Prize laureates in *physics*, *chemistry* and *physiology or medicine* have a sizable fraction of never-cited publications. Although, in the present article, we have not investigated 'average' scientists, we hypothesize that their fraction of never-cited publications is of similar proportions (further research should confirm this hypothesis). In this regard, top scientists do not seem exceptional. A similar observation can be made for other relations between the number of publications, the average number of citations per publication and the h-index. Empirically we found (omitting indices for reasons of simplicity):

$$T \nearrow \implies h \nearrow$$

$$\mu \nearrow \implies h \nearrow$$

$$T \nearrow \implies n \nearrow$$

$$\mu \nearrow \implies n \searrow$$

$$h \nearrow \implies n \nearrow$$

$$\mu \nearrow \implies n/T \searrow$$

where the symbols $\nearrow$ and $\searrow$ stand for strictly increasing, resp. strictly decreasing. The most remarkable result is the fact that when h increases also n increases, which is due to the fact that, empirically, n and h are increasing functions of T. These results are partially confirmed in the theoretical section.

Our results do not explain in a concrete way why Nobel laureates and fields medalists write relatively many publications that are never-cited. We only show that their publication-citation pattern follows the usual distributional lines as e.g. explained by

Lotka's law. Further explanations are needed, which would require the help of experts in the respective fields and/or of Nobel laureates themselves.

In the above empirical study we considered uncited and never-cited publications. In the same way one could study lowly cited publications, for which a threshold on the number of received citations per publication is used. We conjecture that, qualitatively, the same results would have been found (and we modeled this situation in the theoretical part). One could consider it a downside of the present theoretical part that the Lotka distribution lacks the case of 0 citations. It would therefore be interesting to explore how the indicators studied here are interrelated in a framework that is based on a distribution which does include the 0, such as a Lomax or Pareto type 2 distribution. We leave this as an open problem.

Another interesting idea would be to focus on highly cited publications instead of lowly cited ones, and their relation with some of the indicators used here. Of course, one could also study other scientific fields besides the ones we explored.

Finally, it would be interesting to compare the top scientists studied here with 'average' ones, in order to discover how general the phenomena studied are. It seems likely that the number and fraction of never-cited publications of an average scientist is at least as high as those of a Nobel laureate and Fields medalist. Indeed, as recently remarked by Danell (2011), "highly cited authors tend to write the highly cited articles, but all authors can write uncited articles."

References

Q.L. Burrell (2002). Will this paper ever be cited? *Journal of the American Society for Information Science and Technology*, 53(3), 232-235.

Q.L. Burrell (2005). Are "Sleeping Beauties' to be expected? *Scientometrics*, 65(3), 381-389.

R. Danell (2011). Can the quality of scientific work be predicted using information on the author's track record? *Journal of the American Society for Information Science and Technology*, 62(1), 50-60.

L. Egghe (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics*. Oxford (UK): Elsevier.

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

E. Garfield (1980). Premature discovery or delayed recognition – Why? *Current Contents*, #21, p. 5-10, May 26. Available online at http://garfield.library.upenn.edu/essays/v4p488y1979-80.pdf

W. Glänzel, K. Debackere, B. Thijs and A. Schubert (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics, 67(2), 263-277.

W. Glänzel, B. Schlemmer, and B. Thijs (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3), 571-86.

J.E. Hirsch (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.

YX. Liu, I.K. Ravichandra Rao and R. Rousseau (2009). Empirical series of journal h-indices: the JCR category *Horticulture* as a case study. *Scientometrics*, 80(1), 59-74.

A.F.J. van Raan (2004). Sleeping beauties in science. *Scientometrics*, 59(3), 461-466.