View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

Scientometrics (2010) 84:465–479
DOI 10.1007/s11192-009-0098-7

The stability of the *h*-index

Monika Henzinger · Jacob Suñol · Ingmar Weber

Received: 14 August 2009 / Published online: 30 September 2009
© Akadémiai Kiadó, Budapest, Hungary 2009

Abstract Over the last years the *h*-index has gained popularity as a measure for comparing the impact of scientists. We investigate if ranking according to the *h*-index is stable with respect to (i) different choices of citation databases, (ii) normalizing citation counts by the number of authors or by removing self-citations, (iii) small amounts of noise created by randomly removing citations or publications and (iv) small changes in the definition of the index. In experiments for 5,283 computer scientists and 1,354 physicists we show that although the ranking of the *h*-index is stable under most of these changes, it is unstable when different databases are used. Therefore, comparisons based on the *h*-index should only be trusted when the rankings of multiple citation databases agree.

Keywords *h*-index · Ranking scientists · Stability analysis · Citation databases

Introduction

How can one objectively measure the impact of the work of a scientist? This question is of great importance in hiring, promoting and funding decisions in the scientific domain. The work of a scientist is usually defined as his or her set of scientific publications. So measures using the bibliographic record seem natural. Quantities such as the number of publications or the total number of citations these publications have received could be applied. These two quantities, however, have weaknesses as the number of publications merely measures *quantity* and not impact, whereas the total number of citations exclusively measures *impact* but can be inflated by a single publication and does not reward productivity. To address such concerns, in 2005 Hirsch proposed the so-called *h*-index (Hirsch 2005) which has

M. Henzinger · J. Suñol · I. Weber (✉)
Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: ingmar.weber@epfl.ch

M. Henzinger
e-mail: monika.henzinger@epfl.ch

J. Suñol
e-mail: jacob.sunolcanadas@epfl.ch

become the new de-facto standard among such measures (Ball 2007). This index is defined as follows: "A scientist has index h if h of his or her N_P papers have at least h citations each and the other $(N_P - h)$ papers have $\leq h$ citations each" (Hirsch 2005). To justify or question its use in practice, several studies have looked at its predictive power ("Is the h-index correlated with *future* performance?" (Hirsch 2007)), its similarity to other bibliographic measures ("Is a ranking of scientists according to the h-index really different from rankings according to other measures?" (van Raan 2006; Kelly and Jennions 2006)) and its similarity with peer-reviews ("Are quality evaluations by peers correlated to the h-index?" (Bornmann and Daniel 2005)). Similarly, people have commented on its dependence on the age of a scientist (Kelly and Jennions 2006; Kelly and Jennions 2007) and its absolute change when self-citations are disregarded (Schreiber 2007). A recent survey of studies on the h-index can be found in (Bornmann and Daniel 2009). We take a different angle and argue that in order to deserve any credibility an index must be *stable*.

By stable we mean that the ranking induced by the index should change only slightly (i) when the definition of the index itself is changed very slightly, (ii) when a small random set of publications or citations is removed, and (iii) when different bibliographic databases are used. Note that the actual value of any index is irrelevant and that only the corresponding *ranking* matters. Just as the statement that somebody is a millionaire is meaningless unless the prices of goods in the corresponding currency are known, it is meaningless if a person applying for an academic position has a quality measure of 100 unless the corresponding values of other applicants are known. The question of robustness of the h-index under typographic and similar errors was previously studied in (Vanclay 2007) for a single exemplary individual. In that study the author observes a high degree of stability as (i) the relative number of publications affected is small and (ii) errors tend to happen for the long tail of lowly cited publications, which do not affect the h-index.

Our experiments with more than 100,000 cited publications show that when the full range of the h-index is taken into account the h-index is fairly stable under small changes to its definition and under small random perturbations, as might have been introduced by misspellings or by publications missing from the database. Self-citations did not effect the ranking in a noteworthy manner as their total percentage was small but dividing the citations for a publication equally among its authors gave a noticeably different ranking. Most importantly, the ranking showed a big dependency on the database used (ISI Web of Knowledge vs. Google Scholar).

The dataset used and experimental setup

For our experiments we generated a list of 5,614 computer scientists and 1,376 physicists and tried to obtain bibliographic data for them, both from ISI Web of Knowledge (ISI) and from Google Scholar (GS) in January 2009. The computer scientists were a uniform random sample of authors from DBLP,[1] which is the most popular publicly accessible publication database in computer science. The physicists where likewise sampled at random from the physics category at arXiv,[2] which hosts the biggest collection of freely available preprints from physics related domains. When automatically compiling publication lists for authors there exists the problem of multiple scientists sharing the same name. We were careful to avoid this problem and automatically removed author names

[1] http://www.informatik.uni-trier.de/~ley/db/.

[2] http://arxiv.org/.

such as "J. Smith" or "Wen Chan" for which a conflation of different profiles seems likely. Similarly, we removed authors for which we obtained more than one name (e.g., due to marriage) as this would result in an inadequately small publication record (Kurien 2008). Both of these steps were taken to ensure that a scientist in our database corresponds to only one scientist in real life and that, vice versa, a single scientist in real life corresponds to only one scientist in our database. Without these guarantees any conclusions drawn from experiments concerning the stability of the *h*-index could be attributed to the inaccuracy of the database rather than to the *h*-index itself.

We also cleaned the publication data for individual scientists to avoid reporting an editor as author for all her edited contributions (as is incorrectly done by default in ISI) and to remove patents from GS. This cleaning was done as such contributions are generally not considered (i) to be authored by the person under consideration or (ii) to be scientific publications. Furthermore, we tried to ensure a maximal coverage by searching for titles of publications in one database in the other. At the end we manually evaluated the data quality for a sample of scientists to ensure that the publications attributed to a particular person were indeed authored by a single person and that we did not unduly remove legitimate publications. Details of the meticulous data cleaning process and the data quality evaluation can be found in the Appendix. Basic statistics of the data obtained are shown in Table 1.

To quantify the similarity between the ranking according to the original *h*-index and according to variations, we used the following two measures. The first measure is the "percentage of agreeing pairs" *p*. If scientist A is ranked above scientist B in ranking R1, and if this relative order is preserved in R2, then the pair (A,B) is said to be in agreement for R1 and R2. If the relative order is swapped, the pair is in disagreement. The two rankings R1 and R2 are equal if and only if 100% of pairs are in agreement, and one is the exact inversion of the other if and only if 0% of pairs are in agreement. A random

Table 1 Basic statics for data obtained from ISI Web of Knowledge (ISI) and Google Scholar (GS)

Group	Computer scientists		Physicists	
Database	ISI	GS	ISI	GS
Scientists searched	5,614	5,614	1,375	1,375
Scientists found				
… with a publication	4,224	5,163	1,071	1,340
… with a publication in either	5,283	5,283	1,354	1,354
… with a cited publication	2,673	4,461	992	1,243
… with a cited publication in either	4,572	4,572	1,276	1,276
… with $h \in [8,16]$	84	365	243	286
# publications	25,100	59,200	40,800	42,200
# cited publications	11,700	43,100	32,600	25,200
# citations	124,000	649,000	695,000	477,000
# self-citations	3,300	63,200	49,900	54,300
Average *h*-index	2.19	3.54	7.15	6.70
Median *h*-index	1	2	3	4

For our set of computer scientists GS is significantly more comprehensive than ISI, but for the physicists they are comparable. The average and median are computed for the scientists with at least one cited publication in the respective data base. The definition of self-citation is the same that is used by ISI and is explained in a later section

re-ordering will, in expectation, have 50% agreeing pairs. A pair is counted as half-agreement if the two scientists are tied in one ranking but have distinct positions in the other. As a second measure, we looked at how many of the scientists in the top k, say $k = 100$, in ranking R1 are also in the top k in R2. This top k measure quantifies the reliability of the definition of "the super stars" without caring about their relative order. Again, we handled the case where there are ties around position k by allowing fractional overlap between the two sets.

Although these are the only measures for which we report numbers, we also experimented with others. First, the Kendall Tau rank correlation coefficient τ, which is a standard metric to compare two rankings, can be computed from the percentage of agreeing pairs using the formula $\tau = 2p - 1$. Second, we also used the Pearson's correlation coefficient, another standard metric, which led to the same qualitative conclusions as using p or τ. Third, we experimented with a measure which only looks at "significant swaps". The motivation for this is that when using p, the fact that an individual pair is counted as swapped is irrespective of the size of the swap. Even a change from positions (100,101) to (101,100) for the pair (A,B) constitutes a disagreeing pair. Arguably, such small changes should be ignored and only "significant" swaps should be considered. To do this, we also used a variant of the above measure where a pair (A,B) was counted as tied in a ranking unless their ranks differed by at least $n/5$, where n is the number of scientists. Note that if there were two scientists with an equal h-index at, say, the top position, the second position would already have rank 3. Surprisingly, this less sensitive measure gave virtually identical results to the normal version of p, usually with about 1% more agreeing pairs. This indicates that the vast majority of swaps correspond to significant jumps in the hierachy.

In all cases we only computed these similarity measures for the ranking of scientists with at least one cited publication. Scientists without any cited publication will automatically be tied and hence in agreement for all the variations we considered. When we compare the similarity between ISI and GS we consider only scientists which have at least once cited publication in either of the two databases. In addition to the "global" stability analysis for the full spectrum of h-indices, we also analyzed the stability for scientists with an h-index between 8 and 16 (inclusive) in ISI. We chose this range as Hirsch in his original paper suggests "(with large error bars) that for faculty at major research universities, $h \approx 12$ might be a typical value for advancement to tenure (associate professor)" (Hirsch 2005).

As ISI is the more widely used database for studies on the h-index, we only present experimental results concerning the stability for ISI except for the explicit comparison between ISI and GS in one of the later sections. The details of the experimental results for GS, which were qualitatively similar, can be found in the Appendix.

Changes to the definition

Hirsch in his original paper (Hirsch 2005) discussed the advantages and disadvantages of several criteria, such as the total number of papers or the total number of citations, which are commonly used to evaluate scientific output of a researcher. From his discussion one can abstract the following list of "axioms" which any sound measure should satisfy.

1. It must reward productivity.
2. It must reward total impact.
3. It must not be inflated by a small number of "big hits".
4. It must not involve arbitrary thresholds.

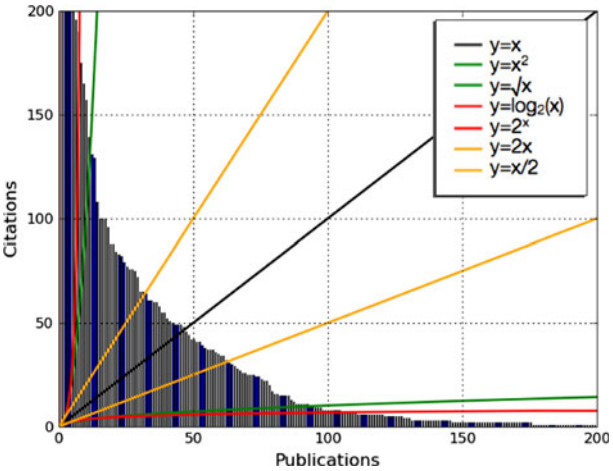

Fig. 1 A citation histogram $c(x)$ for one of the computer scientists in our data set with an h-index of 47. The h-index is defined by the x-coordinate of the intersection between $c(x)$ and the curve $y = x$. Choosing other increasing curves gives more importance to either the initial peak of the curve or to the tail of the curve. Although the actual index changes considerably for different choices, the induced *ranking* is reasonably stable (see online version for colors.)

The h-index does indeed satisfy these axioms. As we will see this advantage applies, however, to a large class of indices.

The h-index can be defined in a different but equivalent way by first sorting the publications of a scientist in descending order of their citation count. Then let $c(x)$ be the number of citations of the x-th most cited publication. By definition $c(x)$ is a non-negative and non-increasing function on the positive real numbers, which reaches $c(x) = 0$ for some finite value of x. For convenience let us assume that $c(x)$ is also continuous. Hirsch's original h-index then corresponds to the x co-ordinate of the point of intersection between the curve $c(x)$ and the curve $f(x) = x$ (or rather the closest integer to the left of x, i.e., $\lfloor x \rfloor$). However, it is easy to see that *any* strictly increasing function $f(x)$ will satisfy the "axioms" listed above. We compared the rankings for the curves $2x, x/2, \sqrt{x}, 2^x, \log_2 x$ and, of course, the original x. Figure 1 shows these different curves and their intersection for the ISI citation histogram of one of the computer scientists in our database. Note that all curves above the original $f(x) = x$ will (i) have a smaller index and (ii) put more emphasis on the height of the initial spike of the curve. On the other hand, curves below $f(x) = x$ will (i) have a larger index and (ii) put more emphasis on the tail of the curve.

If there exists a unique "correct" ranking underlying the h-index then one would expect that these conceptually equivalent indices lead to similar rankings.[3] Our findings partly support this claim. All of the six curves considered agreed for 86–91% and 92–96% of pairs for computer scientists and physicists, respectively. See Table 2 for details. The similarity to the h-index was highest for the "flat" curves $(x/2, \sqrt{x}, \log_2 x)$ and was less pronounced for the "steep" curves $(2x, x^2, 2^x)$. The reason that the steep curves (x^2 and 2^x) are less correlated appears to be that they are more sensitive to a small number of outlier publications with an unusually high number of citations. Having a large number of cited papers appears to be a better indicator for a high h-index than having a small number of highly cited papers.

[3] This is similar to the mathematical notion of "well-definedness" which requires that the choice of a particular representative element from an equivalence class does not effect the outcome.

Table 2 Analysis of the stability of the *h*-index when different curve intersections are used to obtain alternative definitions of the *h*-index as shown in Fig. 1

Range	Similarity measure	$\log_2 x$	$\sqrt{x}$	$x/2$	$2x$	x^2	2^x
$h\in[0,\infty)$	Overlap in top 100	83/86	86/92	86/92	85/93	77/84	76/79
	%-age agreeing pairs	91/96	90/96	90/96	86/95	89/93	85/92
$h\in[8,16]$	Overlap in top 50	44/31	44/38	44/38	42/35	37/28	35/26
	%-age agreeing pairs	82/81	86/87	86/87	82/86	73/76	68/71

The first number in each cell is for computer scientists and the second for physicists

It might be unclear whether to interpret a value such as 80% as indicating that the two rankings are very similar or not. To make such an interpretation possible we suggest the following: The *h*-index was originally introduced to overcome weaknesses of simple measures such as the total number of citations. Arguably, the use of the *h*-index does therefore only make sense if its ranking does indeed differ, as otherwise it still largely coincides with these weaker measures. We therefore consider the similarity between the *h*-index and the total number of citations as a threshold for two indices to be "different". This similarity was computed by ranking scientists according to the total number of citations their publications have attracted and then computing the percentage of agreeing pairs between this ranking and the ranking according to the *h*-index. This percentage turned out to be 79% and 78% for the full and limited *h*-index ranges, respectively, for computer scientists and it was 92% and 82% for physicists. Percentages of agreeing pairs below these reference thresholds are hence considered as "very different". With this convention all of the alternative definitions lead to a similarity above these thresholds, except the versions $\log_2 x$, x^2 and 2^x for the range $h\in[8,16]$.

Normalizing for self-citations and co-author count

Bibliometric measures try to evaluate a scientist's work assuming that recognition by other scientists in the form of citations is a sign of quality or impact. This is similar to how web search engines use hyperlinks and Pagerank to quantify the popularity of a web page. But just as for the internet, where a large number of incoming links created by the owner of the web page invalidate the popularity vote due to a hyperlink, a large number of self-citations can also degrade an impact measure to a mere quantity measure. If, for example, a scientist always cites all of her previous *i*-1 publications in her *i*-th publication, she will after *n* publications have acquired an *h*-index of *n*/2, regardless of any recognition by fellow scientists. As we focused on a per-scientist analysis we only consider a citation to be a self-citation if the author under consideration (co-)authored both relevant publications. So even if publication X cites publication Y and they both have an author in common this does not automatically constitute a self-citation. This is also the definition of self-citation employed by ISI. As the number of such "strict" self-citations was low compared to all citations (see Table 1), removing self-citations had very little impact on the ranking and $\sim$97% of pairs remained in agreement ($\sim$94% for the hiring range). So although the *h*-index is theoretically prone to manipulations by self-citations, this seems to play only a minor role in practice. Part of the reason could be that a publication with an unduly portion of self-citations has little chance of passing a peer-review.

The impact attributed to an individual scientist due to one of her publications is arguably also diminished if it was joint work with a large number of authors. A publication

with 10 authors which is cited 100 times should probably contribute less to each author's *h*-index than a single author paper with 100 citations. The effect of the corresponding normalization, where the number of citations received by a publication is divided by the number of its authors, turned out to be quite noticeable. In the normalized ranking only 82 and 86% of pairs were in agreement for computer science and physics, respectively. In the hiring range the similarity with the *h*-index dropped to 73 and 70%, respectively. As this level of similarity is closer to a *random* ranking than to the ranking or the unmodified *h*-index, it shows that the *h*-index is effected by differences concerning the typical numbers of co-authors. In fact scientists could simply mutually agree to list each other as co-authors on all their publications which, unlike an unduly portion of self-citations, would be unlikely to effect the chances of passing a peer-review process. To avoid this problem we suggest to use the *normalized* variant instead where received citations are divided among authors. The division should, ideally, reflect their contribution to the work or if this is unknown an equal distribution could be used.

Random removal

Any citation database is bound to be incomplete due to different notions of what deserves to be indexed in the first place. Additionally, due to permanent updates a citation database will automatically have a certain amount of inherent "fuzziness" at any point in time.

To test if the *h*-index is sufficiently robust with respect to this type of noise, we analyzed its stability under random removal of (i) citations and (ii) publications. Concretely, in two sets of experiments we removed 10, 25 and 50% of citations or publications uniformly at random. The *h*-index proved surprisingly stable under such perturbations. A removal of 10% of citations (or publications) had very little effect on either set of scientists and at least 95% of pairs remained in agreement. Even after removing 50% of citations (or publications) the agreement was still 88% (81%) and 95% (92%) for computer scientists and physicists, respectively.

Interestingly, such a random removal leads to a *systematic* change in the ranking, and not merely a random change. Concretely, removing 20% of publications uniformly at random has the effect of scaling the function $c(x)$ (see Fig. 1) towards the y-axis by a factor of 0.8. Similarly, removing 20% of citations uniformly at random effectively scales $c(x)$ down towards the x-axis by a factor of 0.8. See Fig. 2 for an illustration of this. A database missing random citations therefore favors left-skewed distributions ("quality" authors),

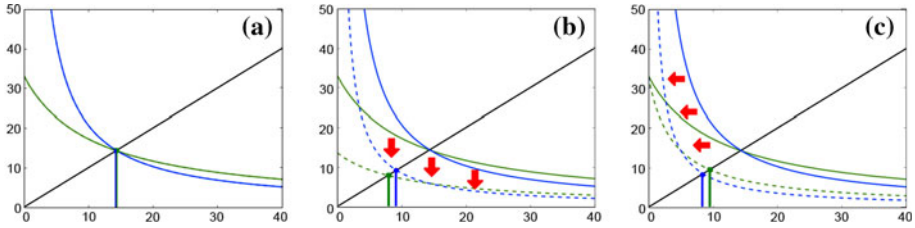

Fig. 2 Removing citations or publications uniformly at random has the effect of scaling the citation histograms towards the x-axis or y-axis, respectively. **a** In the original citation histograms, both the "green" and "blue" scientists have the same *h*-index. **b** Under a random removal of citations the "blue" scientist will have the higher *h*-index. **c** Under a random removal of publications the "green" scientist will have the higher *h*-index (see online version for colors.)

whereas a database missing random publications favors right-skewed distributions ("quantity" authors). This also points towards a dependency on the data collection strategy. For example, a database construction method which is better at extracting citations will lead to a *systematically* different ranking.

Different choice of database

Both of ISI and GS are widely used citation databases and it is unclear which one will be used in any particular instance to compare scientists A and B. As can be seen in Table 1 these two databases vastly differ in their coverage, which also means that the *absolute* values of the *h*-indices for a particular scientist cannot be compared. More fundamental, however, is the question whether both give a similar *ranking* of scientists. This turns out to be not the case.

For all scientists which have at least one cited publication in one of the two databases only 71% of pairs are in agreement for computer scientists and 79% for physicists. For the "hiring range" of scientists with an *h*-index between 8 and 16 in ISI this agreement drops to 67 and 69%, respectively. As an agreement of 50% is achieved by a random reordering, this shows a potential problem with the use of the *h*-index.

One could attribute these differences merely to the difference in coverage (see Table 1). To test whether this is the case, we also computed the *h*-indices using only the subset of publications present in *both* datasets. This leads to a significant increase in the percentage of agreeing pairs to 82 and 92% for computer scientists and physicists, respectively, but remains short of perfect agreement. The only remaining explanations are that the two data bases differ in (i) their total number of citations and/or (ii) their distribution across publications. We explored which of these two was the main reason by trying to "convert" GS to ISI by scaling down (for computer scientists) or up (for physicists) the number of citations appropriately.

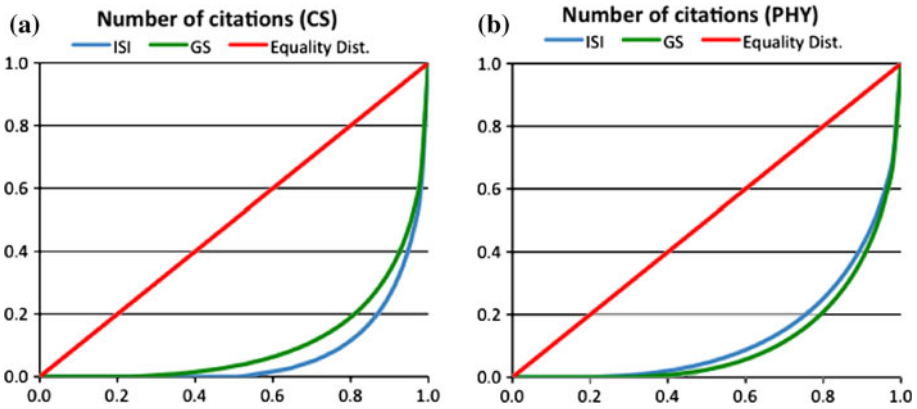

Fig. 3 Google Scholar and ISI differ on how the citations are distributed among the same set of publications. A point (x,y) on a curve means that the least cited x% of the publications have y% of the total citations. The *red line* indicates the situation of perfect equality. For our set of computer scientists GS distributes the citations more evenly than ISI, but for physicists the situation is reversed. **a** Lorentz Curves for computer scientists. **b** Lorentz Curves for physicists (see online version for colors.)

For this scaling we gave every citation in ISI a weight of 1. Then we scaled up/down the number of citations for each publication in GS such that the total weight of citations in GS equals the total weight in ISI. Only publications present in both datasets were considered for this. The corresponding scaling factors were 0.38 for computer scientists and 1.31 for physicists. Surprisingly, this scaling did not have a noticeable additional effect compared to only requiring the set of publications to be identical. This indicates that the distribution of citations across publications differs more fundamentally. To visualize this differences Fig. 3 shows the Lorentz Curves for the two databases and the two sets of scientists. As the scaling will leave the Lorentz Curves *unchanged* these differences remain.

One possible explanation for the different distributions is a difference in the citation collection strategy. One database could care more about depth of citation coverage and try to obtain the complete set of citations for already highly cited publications, while caring less about less cited ones, whereas another data base could emphasize breadth of coverage and try to obtain at least a few citations for all publications, while caring less about obtaining the complete set for highly cited ones.

Conclusions

As the ranking according to the *h*-index differs considerably across different databases we recommend that wherever possible at least two different databases should be consulted and the relative ranking should only be trusted if it is consistent between the databases. Additionally, in some fields it might be advisable to adapt the *h*-index by normalizing each citation count by the number of authors.

Appendix

This appendix serves the following purposes. First, to give details of how the data from ISI Web of Knowledge (ISI) and from Google Scholar (GS) was preprocessed in an attempt to ensure the highest possible data quality. Second, to report the results for the manual evaluation of the data quality. And, finally, to report detailed numbers for the experiments with GS. These three items correspond to the following three sections.

Data preprocessing

One of the biggest problems in the area of automated bibliometrics is the issue of author disambiguation. How can one know that the author "John Smith" of publication A is the same entity as the author "John Smith" of publication B? Though several approaches to address this problem exist, this hard problem remains largely unsolved. As a consequence, studies either only involve a small set of hand-selected scientists whose publication record is then manually checked, or studies trust the online database and simply assume that different names or different combinations of initial and last name correspond to different scientists. We tried to sidestep the disambiguation problem by identifying a set of authors whose names most likely correspond to only one scientist in the subject area of consideration. In the following, we explain in detail how we obtained our final list of both computer scientists and physicists to use for our experiments.

1. *Obtain an initial list of candidate names.* For computer science we used all the 681,000 authors listed in the DBLP database[4] and for physics we used all the 82,000 authors listed under the "Physics" section of the arXiv collection[5] as of January 2009. Only scientists for which we could obtain their *full* first and last name were kept.
2. *Map all names to ASCII characters.* The encoding of the names was converted to ASCII wherever possible using the "Normalization Form Compatibility Decomposition".[6] This included conversions such as "Schrödinger → Schrodinger" and "Suñol → Sunol". Names with characters where this was not possible were removed in this process. Note that GS also returns documents containing "Schrödinger" for the query "Schrodinger" but not vice versa. In ISI the query "Schrödinger" gives an error message and the conversion to "Schrodinger" is required.
3. *Remove common US and Chinese family names.* We used dictionaries of common US and Chinese names and removed corresponding cases such as "Smith" or "Chiang". These cases are especially likely to be ambiguous.
4. *Remove last names with less than five characters.* Very short names such as "Mata" or "Tsai" were removed. This filter removes additional names of Asian origins, for which disambiguation is a serious problem, and it also removes some artifacts where a sequence of initials was contracted to give a "name".
5. *Require unique initial plus last name combination.* All names were mapped to a single initial plus a last name, e.g., "John A. Smith → J. Smith". Parts such as "van" or "de" were treated as part of the last name. For each combination of initial and last name the original list was checked for uniqueness. Only cases with a *unique* combination were maintained. This was required as GS indexes some and ISI all of its publications using only this combination, so that "John Smith" and "Jack Smith" could not be differentiated. As only the first initial was used this filtering step also avoids that "J. A. Smith" is accidentally conflated with "J. B. Smith".
6. *Remove authors with several aliases.* For computer scientists we removed all cases where a single DBLP entity had several aliases. This was done to avoid missing some publications due to a name change of the scientist, e.g., due to marriage. For physicists we did not have a list of such aliases.
7. *Require unique entry in ISI.* When searching for the initial and last name in ISI's Author Finder[7] we required a *unique* matching entry and all other names were discarded.

The final numbers, 5,614 computer scientists and 1,375 physicists, of scientists with at least one publication in either database were obtained by sampling 10,000 and 3,000 scientists, respectively, right before the very last filtering step. Scientists who did not have a unique ISI entry or for whom no publication with the corresponding topic filter (see below) could be found ended up without any publication found and were ignored for all the experiments. After the two lists of names had been compiled, we obtained bibliographic information by querying GS and ISI as follows.

1. *Query GS with full name and subject area filter.* For each scientist we sent the corresponding query with the full first and last name, but without any additional

[4] http://www.informatik.uni-trier.de/ley/db/.

[5] http://arxiv.org/.

[6] http://en.wikipedia.org/wiki/Unicode\s\do5(n)ormalization.

[7] http://apps.isiknowledge.com/OutboundService.do?action=go&mode=afServiceproduct=WOS.

initials, to GS. Here we used Google's topic filter[8] with the "Engineering, Computer Science, and Mathematics" and the "Physics, Astronomy, and Planetary Science" option for computer science and physics, respectively. Topic filters were used to avoid possible name collisions with scientists in other domains. The returned publications were added to a list of candidate publications for this scientist.

2. *Query ISI with initial plus last name and subject area filter.* For each scientist we sent the corresponding query with the initial and last name to ISI, as ISI does not index the full first name. Only publications from a subject related to computer science (or mathematics) or physics were considered. The returned publications were added to a list of candidate publications for this scientist.
3. *Re-Query ISI with titles from GS.* For each item in the current list of GS publications we queried ISI directly for the corresponding title if we had not yet found this publication inside ISI. The returned publications were added to a list of candidate publications for this scientist.
4. *Re-Query GS with titles from ISI.* For each publication found in ISI but not found in Google Scholar we queried GS directly for the corresponding title. If this title was found it was added to a list of candidate publications.
5. *Remove invalid publications.* We removed all patents from GS. Additionally, we made sure that each publication did indeed match the corresponding query as in some cases a query "John Smith" could return a publication with the two authors "John Doe" and "Anne Smith".
6. *Remove duplicates in GS.* Google Scholar sometimes returns the same publication as a duplicate multiple times. In cases where two or more publications with the same title and the same year of publication were found, only the entry with the highest number of citations was kept. To compare titles for equality all non-alphabetic characters were removed before the comparison.

Basic statistics about the database constructed in this way can be found in Table 1. After obtaining all the bibliographic information we manually checked the quality of the data as discussed in the next section.

Quality evaluation

After obtaining the data from ISI and GS as explained in the previous section, we manually evaluated the quality of our database, both for ISI and for GS. Concretely, for 15 computer scientists and 15 physicists we wanted to know the following.

- Unique identity: Are the scientists in our database "unique", i.e., do all the corresponding publications we have correspond to the same person?
- Data quality: Are the publications we have "valid" publications, e.g., are they authored by the corresponding scientist, do they belong to the correct topic (computer science or physics) and are they not duplicates?
- Data coverage: Do we have *all* valid publications for a scientist, i.e., did we not miss relevant publications pertaining to the person under consideration?

[8] http://scholar.google.com/advanced\s\do5(s)cholar\s\do5(s)earch.

Unique identity

To determine if different publications stored under the same author name in our database actually belong to the same person, we looked at clues from the affiliation, the co-authors and the topics of the publications. We considered two publications to belong to two different authors if the affiliations differed, if the sets of co-authors are disjoint and if the topics of the publications appear to be unrelated. For this test we only looked at publications with at least one citation in either ISI or GS as only these publications are relevant for our experiments. Among the 15 computer scientists checked, *all* were indeed unique and all their publications belonged to a single individual, both in ISI and in GS. For the physicists, there was one case where only three out of four cited publications in GS were authored by the same person and one publication belonged to an individual with the same combination of first name and last name. This person was, however, unique in ISI.

Data quality and coverage

To determine both the quality or cleanness, referred to as *precision*, and coverage or completeness, referred to as *recall*, of our database, for each scientist we tried to compile a ground truth list of all the person's publications. In some cases, such a list could be found on a homepage. In other cases it was impossible to do better than combining and filtering manually the publications for a general query by initial and last name in both ISI and GS. Publications which were clearly not on the topic under consideration (either computer science or physics) were removed from the list.

Once this list was constructed, we then compared the publications in our database to the entries on the list. To reduce the work required for the quality test we only checked the validity of the publications with at least as many citations as the person's h-index. Other, rarely cited publications, do not contribute to the computation of the h-index. For the test of completeness we tried to find all list entries in our database, regardless of whether they were cited or not.

The focus of the quality evaluation was to show that our preprocessing actually improves the quality of the data compared to, say, simply querying for a person's last name and initial without any additional filters. Put differently, if a valid publication cannot be found at all in ISI, even if searching for (parts of) its title directly, then no data preprocessing in the world will manage to extract information about this publication from ISI. This is then not a shortcoming of the preprocessing but of the underlying dataset. On the other hand, if a valid publication is lost in the preprocessing then the filtering process might be too aggressive.

Table 3 compares the precision and recall of our extracted and cleaned ISI and GS database with the following simpler versions of obtaining the data. Concerning ISI one could consider the two options of (i) only searching for the initial and last name of a person or (ii) also making use of ISI's topic filter. As for Google Scholar one could (i) search for the full name of a person, (ii) search for the full name in combination with a topic filter, (iii) search for the initial and last name, or (iv) search for the initial and last name with a topic filter. All of these alternatives are listed in Table 3. Comparing the results for the extracted versions of ISI and GS with the results for their "raw" counterparts shows that our data preprocessing does indeed help to improve the precision without sacrificing too much recall. For example, for Google Scholar the precision in our extracted database is 0.94, which could also have been achieved by searching for the full name with a topic filter directly. However, our recall is still over 0.80, compared to less than 0.60 when searching only for the full name with a topic filter.

Table 3 "Initial + last name" vs. "full name" indicates how the raw database was queried

	Extracted database		"Raw" ISI		"Raw" Google Scholar			
	Our preprocess		Initial + last name		Initial + last name		Full name	
	ISI	GS	w/o TF	w/TF	w/o TF	w/TF	w/o TF	w/TF
Precision	0.92/0.92	0.94/0.94	0.53/0.79	0.92/0.93	0.42/0.66	0.62/0.82	0.72/0.87	0.94/0.94
Recall	0.63/0.73	0.85/0.81	0.65/0.75	0.63/0.72	0.94/0.88	0.75/0.77	0.70/0.52	0.53/0.41

The first option has a higher recall, as more author names will match, whereas the second has a higher precision. TF stands for topic filter and indicates that the corresponding option in ISI/GS was used. Again, using this filter improves the precision at the expense of recall. Precision and recall values are computed on a per-person basis and then averaged over all 15 computer scientists (first number in each cell) or physicists (second number in each cell), respectively

Results for GS

In the main publication we only reported numbers for ISI, except for the explicit similarity comparison with GS. In this section we also give the results for GS. For easier comparison purposes, we also include the results for ISI again.

Changes to the definition

As mentioned in the main article, the *h*-index can be viewed as a member of a large family of possible indices, which all satisfy certain axioms. If all these variants give similar rankings, then this is an indication that the observed ranking is indeed a property of the data rather than a property of the particular index.

To put the percentage of agreeing pairs into perspective it is again helpful to compare with the similarity between the ranking according to the *h*-index and the ranking according to the total number of citations, as the *h*-index aims to be a different index. For Google Scholar the corresponding percentages of agreeing pairs were 85 and 90% for all computer scientists and computer scientists with an *h*-index between 8 and 16, respectively. The corresponding numbers of physicists were 91 and 86%. Similarity values below these threshold can, arguably, be seen as indicating substantially *different* rankings. Table 4 shows that for Google Scholar as for ISI (see main article) the *h*-index is stable under most changes, though usage of the curves x^2 and 2^x leads to noticeably different rankings.

Table 4 Analysis of the stability of the *h*-index when different curve intersections are used to obtain alternative definitions of the *h*-index

Database	Range	Similarity measure	$\log_2^x$	$\sqrt{x}$	$x/2$	$2x$	x^2	2^x
ISI	$h\in[0,\infty)$	Overlap in top 100	83/86	86/92	86/92	85/93	77/84	76/79
		%-age agreeing pairs	91/96	90/96	90/96	86/95	89/93	85/92
	$h\in[8,16]$	Overlap in top 50	44/31	44/38	44/38	42/35	37/28	35/26
		%-age agreeing pairs	82/81	86/87	86/87	82/86	73/76	68/71
GS	$h\in[0,\infty)$	Overlap in top 100	83/87	87/91	87/91	90/91	74/77	67/68
		%-age agreeing pairs	93/94	92/95	92/95	90/94	90/91	88/91
	$h\in[8,16]$	Overlap in top 50	46/37	40/43	48/37	48/42	44/33	42/35
		%-age agreeing pairs	89/85	94/90	94/90	94/89	87/82	83/80

The first number in each cell is for computer scientists and the second for physicists

Normalizing for self-citations and co-author count

Table 5 shows the stability of the *h*-index when (i) self-citations are removed and when (ii) citations are normalized by the number of authors. As discussed in the main article, the ranking changes very little when self-citations are removed as the percentage of "strict" self-citations is very small. This hold both for ISI and for GS. However, normalizing the number of citations by the number of authors does have a noticeably impact. Here, for the

Table 5 Analysis of the stability of the *h*-index when (i) self-citations are removed and (ii) a publication's citation count is normalized by the number of its authors

Database	Range	Similarity measure	w/o self-citations	w/author normaliz.
ISI	$h\in[0,\infty)$	Overlap in top 100	91/96	77/67
		%-age agreeing pairs	99/98	82/86
	$h\in[8,16]$	Overlap in top 50	46/42	39/33
		%-age agreeing pairs	92/93	73/70
GS	$h\in[0,\infty)$	Overlap in top 100	93/96	80/80
		%-age agreeing pairs	95/97	86/88
	$h\in[8,16]$	Overlap in top 50	48/46	45/36
		%-age agreeing pairs	95/94	88/80

The first number in each cell is for computer scientists and the second for physicists

Table 6 The percentage indicates the fraction of (a) citations and (b) publications removed

Database	Range	Similarity measure	10%	25%	50%	95%
(a) Random removal of citations						
ISI	$h\in[0,\infty)$	Overlap in top 100	92/96	92/95	88/93	56/80
		%-age agreeing pairs	97/99	94/97	88/95	79/86
	$h\in[8,16]$	Overlap in top 50	46/43	45/41	42/36	35/23
		%-age agreeing pairs	92/94	88/91	83/86	69/70
GS	$h\in[0,\infty)$	Overlap in top 100	94/97	93/96	89/92	68/73
		%-age agreeing pairs	98/98	95/97	91/94	80/84
	$h\in[8,16]$	Overlap in top 50	47/46	47/45	44/42	33/33
		%-age agreeing pairs	98/96	96/93	94/89	84/75
(b) Random removal of publications						
ISI	$h\in[0,\infty)$	Overlap in top 100	91/94	86/93	78/89	41/72
		%-age agreeing pairs	95/98	88/95	81/92	73/77
	$h\in[8,16]$	Overlap in top 50	45/42	43/38	40/33	33/17
		%-age agreeing pairs	90/92	84/86	78/79	63/62
GS	$h\in[0,\infty)$	Overlap in top 100	91/94	89/91	83/87	56/63
		%-age agreeing pairs	96/97	91/94	85/90	71/73
	$h\in[8,16]$	Overlap in top 50	45/45	43/42	42/39	27/28
		%-age agreeing pairs	96/94	93/89	90/82	72/62

For the reduced dataset the ranking of the *h*-index is then compared to the unperturbed setting. All numbers are averaged over 50 runs, where one run corresponds to a concrete random instance of removed citations/publications. The first number in each cell is for computer scientists and the second for physicists

range $h \in [8,16]$ the similarity drops below the similarity with the ranking according to the total number of citations.

Random removal

Table 6 shows the stability of the *h*-index when citations or publications are removed uniformly at random. Both for ISI and for GS the *h*-index is very robust under such perturbations.

References

Ball, P. (2007). Achievement index climbs the ranks. *Nature, 448*(7155), 737.
Bornmann, L., & Daniel, H. (2005). Does the h-index for ranking of scientists really work? *Scientometrics, 65*(3), 391–392.
Bornmann, L., & Daniel, H. D. (2009). The state of h index research. Is the h index the ideal way to measure research performance? *EMBO reports, 10*(1), 2–6.
Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572.
Hirsch, J. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences, 104*(49), 19193.
Kelly, C., & Jennions, M. (2006). The h index and career assessment by numbers. *Trends in Ecology & Evolution, 21*(4), 167–170.
Kelly, C., & Jennions, M. (2007). H-index: age and sex make it unreliable. *Nature, 449*(7161), 403.
Kurien, B. (2008). Name variations can hit citation rankings. *Nature, 453*(7194), 450.
Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Europhysics Letters, 78*(3), 0295–5075.
van Raan, A. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502.
Vanclay, J. K. (2007). On the robustness of the h-index. *Journal of the American Society for Information Science and Technology, 58*, 1547–1550.