PLOS ONE

RESEARCH ARTICLE

A New Bibliometric Index Based on the Shape of the Citation Distribution

Tommaso Lando[1], **Lucio Bertoli-Barsotti**[2]*

1. Department of Finance, VŠB Technical University of Ostrava, Ostrava, Czech Republic, **2.** Dipartimento di Scienze aziendali, economiche e metodi quantitativi, University of Bergamo, Bergamo, Italy

*lucio.bertoli-barsotti@unibg.it

Abstract

In order to improve the *h*-index in terms of its accuracy and sensitivity to the form of the citation distribution, we propose the new bibliometric index *l*. The basic idea is to define, for any author with a given number of citations, an "ideal" citation distribution which represents a benchmark in terms of number of papers and number of citations per publication, and to obtain an index which increases its value when the real citation distribution approaches its ideal form. The method is very general because the ideal distribution can be defined differently according to the main objective of the index. In this paper we propose to define it by a "squared-form" distribution: this is consistent with many popular bibliometric indices, which reach their maximum value when the distribution is basically a "square". This approach generally rewards the more regular and reliable researchers, and it seems to be especially suitable for dealing with common situations such as applications for academic positions. To show the advantages of the *l*-index some mathematical properties are proved and an application to real data is proposed.

Citation: Lando T, Bertoli-Barsotti L (2014) A New Bibliometric Index Based on the Shape of the Citation Distribution. PLoS ONE 9(12): e115962. doi:10.1371/journal.pone.0115962

Editor: Lutz Bornmann, Max Planck Society, Germany

Received: September 19, 2014

Accepted: December 2, 2014

Published: December 26, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are within the paper.

Funding: TL's research has been elaborated in the framework of the project Opportunity for young researchers, reg. no. CZ.1.07/2.3.00/30.0016, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic (www.msmt.cz). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The main success of the *h*-index [1] is probably due to its simplicity and its robustness, in that it is insensitive to low-impact publications with few or no citations. On the other hand, the drawbacks of the *h*-index have been discussed. Due to its symmetric structure [2], the *h*-index is insensitive to highly-cited publications: as soon as one such publication is part of the *h*-core (the group of the *h* most highly-cited papers; [3]), its actual number of citations no longer has an influence. Moreover, the number *h* alone seems to be too poor to discriminate among authors with similar scientific productions. This problem is known as the

"low resolution" [4] of the Hirsch index: indeed, it is quite common to find researchers with equal h values. For these various reasons, several methods to complement or to improve the h-index have been proposed. The A-index [5], the R-index [6] and the e-index [4] complement the h-index by measuring the overall citation "intensity" in the h-core. On the other hand, the main stand-alone alternative to the h-index is probably the g-index [7], which is sensitive to exceptional publications, although it is not really sensitive to the form of the citation distribution. Other h-type indices attempt to improve the h-index by extracting additional information from the form of the citation distribution. We list some of these alternative approaches: the tapered h-index (h_T, [8]); the Zynergy index (z-index) [9]; the recently introduced h'-index [10].

For a given author x, let $\mathbf{x}$ be his/her corresponding citation distribution - that is, the vector of non-negative integer components representing the number of citations per publication (as usual, in this paper we will assume that the citation distribution is sorted in decreasing order) - and let C_x be the total number of citations. Our idea is to propose a new bibiometric index which depends on the *similarity* between the citation distribution $\mathbf{x}$ and a corresponding "ideal" distribution $\mathbf{x}^*$, to be uniquely identified, under suitable constraints, in terms of i) number of papers; ii) number of citations per publication. More precisely, we search for an index which increases its value as the citation distribution $\mathbf{x}$ approaches the ideal form defined by $\mathbf{x}^*$. For instance, we could possibly define $\mathbf{x}^*$ as a distribution with a "rectangular" form (henceforth we use this term to denote a vertical rectangle, i.e. most of the citations are "concentrated" on one or a few papers). This approach would reward researchers with a high impact on the scientific community (rather than regular productivity) and might be appropriate if it is necessary to evaluate high-level scientists (e.g. Nobel-prize winners or Fields medalists), but it could be misleading in many common contexts. In this paper we shall not follow this logic. In fact, study of the Hirsch index and its most important alternatives shows that the scientific performance of an author is always maximized if the distribution is basically represented by a "square" with side $\left[\sqrt{C_x}\right]$ (where $[t]$ is the integer part of the number t): in this case we find that h and g (as shown in the next section) both reach their maximum values, as well as other bibliometric indicators. For this reason, in this paper we choose to define $\mathbf{x}^*$ on the basis of a "squared" form. This idea yields a bibliometric index which is especially suitable for evaluating the scientific performance of "standard level" researchers. Consider the common case when the evaluation of a researcher is intended to assess his/her suitability for an academic position, e.g. as full professor etc. We believe that in such situations bibliometric indicators are especially useful. If applicants are similar/comparable, we believe that a bibliometric index should reward the more regular researchers in order to enable research institutions to make reliable selections. Thus, in a bibliometric context, a sort of "risk-averse" attitude suggests choosing, between researchers of the "same level" (that is, with equal or similar number of citations), the one who produces a good number of

good quality papers, and who therefore has a more regular (i.e. ''squared'') distribution of citations.

Although a general class of indices is proposed, we subsequently focus on a particular index, defined as l. The mathematical properties of l are presented and formally proved: l is a novel bibliometric indicator which outperforms the h-index in terms of accuracy and sensitivity to the form of the citation distribution. An application to real data shows that l is strongly correlated with other important h-type indices. Moreover, we attempt to analyze the dependence between bibliometric rankings and the judgements of a committee, obtaining interesting results for the new index l.

Methods

For a given researcher x with a total number of publications n_x let us denote with x_i the number of citations of paper i ($i=1,\ldots,n_x$), and let the papers be ranked in decreasing order according to the number of citations that they have received, so that $x_1 \geq x_2 \geq \ldots \geq x_{n_x}$. Let us denote the vector $\mathbf{x}=(x_1,x_2,\ldots,x_{n_x})$ by the *citation distribution*. Henceforth let us call *a-core* (for any positive integer a) the set of the a most cited papers (if it exists). A bibliometric index of author x is a mathematical function of his/her citation distribution $\mathbf{x}$.

The h-index [1] is defined as follows:

$$h(\mathbf{x}) = max\{i \mid x_i \geq i\}. \tag{1}$$

The number h identifies a set of significant papers, the so-called *h-core*. It is interesting to observe that the Hirsch index mainly depends on the form of the citation distribution: h is greater when the distribution is ''squared'' and smaller when the distribution has a ''rectangular'' form. In particular, $h(\mathbf{x})$ cannot exceed $\overline{m}(\mathbf{x}) = \min\{x_1,\overline{n}_x,\lceil\sqrt{C_x}\rceil\}$ where $\overline{n}_x$ is the number of papers with at least 1 citation [11]. *A fortiori*, for any author x with a fixed number of total citations $C_x = C$ the value of $h(\mathbf{x})$ cannot exceed $\lceil\sqrt{C}\rceil$. In particular, the distribution $\mathbf{s}=(s_1,s_2,\ldots,s_{n_x})$, $s_i \in \mathbb{N}$, with total citations C and such that $s_i \geq \lceil\sqrt{C}\rceil$ for $i=1,\ldots,\lceil\sqrt{C}\rceil$ yields $h(\mathbf{s}) = \lceil\sqrt{C}\rceil$. Note that $\mathbf{s}$ can be basically represented by a ''square'' with side $\lceil\sqrt{C}\rceil$. To be more specific, we can say that, for any possible citation distribution $\mathbf{x}$ such that $C_x = C$:

$$\max_{\mathbf{x}} \; h(\mathbf{x}) = h(\mathbf{s}) = \left\lceil \sqrt{C} \right\rceil . \tag{2}$$

One of the main alternatives to the h-index is the g-index, proposed by Egghe [7]. The g-index is defined as:

$$g(\mathbf{x}) = max\left\{i \mid X_i \geq i^2\right\}, \tag{3}$$

where $X_i = \sum_{j=1}^{i} x_j$. Similarly to h, the number g identifies a set of significant papers, the *g-core* (note that this set may be constituted by fictitious publications without citations, when $n < g$; [12]). It is interesting to note that $x_i \geq i$ for $i = 1, \ldots, h(\mathbf{x})$ yields $X_i/i \geq i$ for $i = 1, \ldots, h(\mathbf{x})$; thus, by definition, the h-core is a subset of the g-core ($g \geq h$, as is well known). The g-index is sensitive to highly-cited publications and does not strictly depend on the form of the distribution. Indeed it is known that g is sensitive to *concentrative transfers* [13], [12]. Hence, for a given number of total citations C, a distribution which concentrates all these citations on a single paper maximizes g. Actually, unlike h, the g-index can be maximized by both a "squared" and a "rectangular" distribution: from this point of view we can say that the g-index is more "flexible" than the h-index. On the other hand, this shows that g does not depend on the form of the distribution. This result can be proved as follows. Define by $I(P)$ the logical function such that $I(P) = 1$ if the proposition P holds true and $I(P) = 0$ otherwise. For any author x with $C_x = C$ citations, consider the corresponding "rectangular" distribution: $\mathbf{r} = (C, 0, \ldots, 0)$ (vector with n_x elements, for instance). Observe that:

$$\max_{\mathbf{x}} g(\mathbf{x}) = g(\mathbf{r}) = g(C, 0, \ldots, 0) = \sum_i I\left(\left\lceil \sqrt{C} \right\rceil \geq i\right) = \left\lceil \sqrt{C} \right\rceil. \tag{4}$$

Let $\mathbf{s} = (s_1, s_2, \ldots, s_{n_x})$ be the "squared form" distribution such that $s_i \geq \lceil \sqrt{C} \rceil$ for $i = 1, \ldots, \lceil \sqrt{C} \rceil$ ($\mathbf{s}$ can be obtained from $\mathbf{r}$ by a finite number of *elementary transfers*, called *T-transforms* in [14, p.32]. Consider that, for $\mathbf{s}$, we obtain $S_i = \sum_{j=1}^{i} s_j \geq i\lceil \sqrt{C} \rceil$; thus

$$g(\mathbf{s}) \geq \sum_i I(i\left\lceil \sqrt{C} \right\rceil \geq i^2) = \sum_i I\left(\left\lceil \sqrt{C} \right\rceil \geq i\right) = \left\lceil \sqrt{C} \right\rceil, \tag{5}$$

hence $g(\mathbf{s}) = \lceil \sqrt{C} \rceil$. We conclude that $\max_{\mathbf{x}} g(\mathbf{x}) = g(\mathbf{s}) = g(\mathbf{r}) = h(\mathbf{s}) = \max_{\mathbf{x}} h(\mathbf{x}) = \lceil \sqrt{C} \rceil$. Note that this results can also be derived from the bounds of the h- and g-indices recently studied by [15].

Overall, it seems that both indices (h and g) agree when the citation distribution is squared, which happens when a researcher produces a significant number of good quality publications, rather than a few outstanding ones. As a consequence of this idea, which is apparently consistent with the most popular bibliometric indices, we propose to measure the scientific performance of a researcher by comparing his/her citation distribution to a squared benchmark distribution, as described in the next subsection.

Defining an "ideal" citation distribution

Define $h^*(\mathbf{x}) = \lceil \sqrt{C_x} \rceil$. The number h^* corresponds to a set of papers which includes the h-core as well as the g-core. It is worth noting that it may happen that an author does not have h^* published papers (i.e. when $h^*(\mathbf{x}) > n_x$ which is quite uncommon, especially for "standard" researchers): we may consider h^* as an

"ideal" number of papers. If author x with C_x citations has at least $h^*(\mathbf{x})$ publications, then (according to the citation distribution) he/she can maximize his/her scientific performance (in terms of both h and g); otherwise he/she cannot. In the literature, several methods have been proposed to select the optimal number of significant or "elite" papers which have a high impact on the scientific community. Generally, bibliometric indicators based on larger sets are more appropriate to measure the overall performance instead of scientific impact. On the other hand, indices that focus on a smaller set or "core" of highly cited papers assess authors based on their impact, overlooking the regularity of their performance. The $P_{top10\%}$ index [16, 17] is the number of papers which belong among the top 10% highly cited publications on the same subject and in the same year; obviously by varying the percentage we can obtain more or less restricted elite sets. One of the main advantages of this approach is that it makes it possible to compare authors in different research fields and different periods of time. Nevertheless, the aim of the $P_{top10\%}$ index is quite different from ours, and we do not have available the data for its computation; for these reasons the $P_{top10\%}$ is not included in our analysis. The π-index [18, 19] is obtained from the citations within the π-core, that is, the set of the most $\left\lceil\sqrt{n_x}\right\rceil$ cited papers. Generally, the π-index considers the most elite papers and therefore rewards papers of high impact, although the π-core depends on the number of publications, which is not a measure of impact itself. Moreover, other indicators such as the above mentioned A-, R- and e-indices are based on the number of citations within a generally larger set i.e. the h-core. Note that these indices have been proposed as complementary to h and not as "stand-alone" indicators due to some possible drawbacks (e.g. an increase in h could produce a decrease in A or e). The aim of this paper is to take into account not only the impact but also the regularity of an author during his/her entire career. In fact, as mentioned above and confirmed by our case study, we are interested in assessing "standard level" researchers who possibly do not have outstandingly higlhy cited papers. We therefore propose to consider the h^*-core, which generally includes the h-core, as well as the π-core.

As discussed above, the h- and g-indices can be maximized by a "squared" citation distribution (with side equal to h^*). It is worth noting that, for a fixed number of citations C_x, a distribution of this kind also maximizes other alternative h-type indices, such as the h_T-index [8] and the R-index [6]. Therefore, some of the most important bibliometric indices suggest that a "squared-form" citation distribution should represent an "ideal" for an author. Also, the z-index [9] complies with this principle, because z increases with consistency (regularity, see [20]). We have maximum consistency in the case of absolutely uniform performance [24], that is, when all the papers have an equal number of citations. We believe that the best performance can be achieved when a combination of impact (citations per paper), productivity (number of papers) and consistency is maximized, and this happens with a "squared" distribution. In particular, we propose to define an ideal number of citations per paper as described below.

Assume that author x has at least one publication and one citation. Define R_x as the natural number such that $(h^*(\mathbf{x}))^2 + R_x = C_x$. Given n_x, C_x and R_x we can now define an ideal citation distribution, say $\mathbf{x}^*$, such that $h(\mathbf{x}^*) = g(\mathbf{x}^*) = h^*(\mathbf{x}^*)$. Although there may be different (also easier) ways to define $\mathbf{x}^*$, we propose choosing the distribution $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_{h^*(\mathbf{x})+1}^*)$ (a vector with h^*+1 components) which reflects maximal regularity, in that $x_i^* \geq h^*(\mathbf{x})$ as long as possible (for $i = 1, \ldots, h^*(\mathbf{x})$) and R_x is symmetrically equidistributed among papers/citations. This idea is formalized as follows:

$$x_i^* = \begin{cases} h^*(\mathbf{x}) + I(i \leq \frac{R_x+1}{2}), & \text{for } i = 1, \ldots, h^*(\mathbf{x}), \\ \sum_{j=1}^{h^*(\mathbf{x})} I(j < \frac{R_x+1}{2}), & \text{for } i = h^*(\mathbf{x})+1. \end{cases} \quad (6)$$

Thus, the components of $\mathbf{x}^*$ are all positive integer numbers except for the last one ($x_{h^*(\mathbf{x})+1}^*$), which can possibly be 0. The choice of a vector $\mathbf{x}^*$ with $h^*(\mathbf{x})+1$ components instead of $h^*(\mathbf{x})$ is due to the fact that, with this choice, we can "distribute" R_x in the most efficient way in order to maximize the most important bibliometric indices. Let the symbol "$\cong$" represent a generalized equality between vectors which simply excludes the zero-elements from $\mathbf{x}$ ($\mathbf{x} \cong \mathbf{x}^*$ if $\overline{\mathbf{x}} = \overline{\mathbf{x}}^*$, where for a k-dimensional citation vector $\mathbf{a}$, define $\overline{\mathbf{a}} = (a_1, a_2, \ldots, a_{\overline{k}})$ and $\overline{k} = \#\{a_i \neq 0, i = 1, .., k\}$). Note that the citation distribution defined by $\mathbf{x}^*$ maximizes h, g and also the h_T-index [8], so that it is evident that any researcher x for whom $\mathbf{x} \cong \mathbf{x}^*$ really optimizes his/her scientific performance.

A bibliometric index based on the form of the citation distribution

For any author x, it is now possible to obtain a class of bibliometric indices which are sensitive to the similarity between the real distribution $\mathbf{x} = (x_1, x_2, \ldots, x_{n_x})$ and the corresponding ideal distribution $\mathbf{x}^*$. The basic idea is that, between two scientists x and y of the same level, i.e. with the same number of total citations $C_x = C_y = C$, the one (say x) whose distribution $\mathbf{x}$ is more "similar" to $\mathbf{x}^* = \mathbf{y}^*$ should be preferred (it is easier for author x to reach his/her maximum h- and g-values $h^*(\mathbf{x})$ compared to y).

Denote by n_x^* the number of papers such that $x_i^* > 0$ (n_x^* can be equal to $h^*(\mathbf{x})$ or to $h^*(\mathbf{x})+1$ depending on R_x) and assume that, in the rare case when $n_x < n_x^*$, $x_j := 0$ for $j = n_x, \ldots, n_x^*$. Drawing inspiration from statistical divergence measures between distributions [21], we can measure the "distance" between $\mathbf{x}$ and $\mathbf{x}^*$ by analyzing the ratios $\frac{x_i}{x_i^*}$, for $i = 1, .., n_x^*$: if they are (on average) close to 1, we can conclude that $\mathbf{x}$ is close to $\mathbf{x}^*$. Suppose that the citation distributions $\mathbf{x}$ and $\mathbf{y}$ yield the same ideal distribution $\mathbf{x}^* = \mathbf{y}^*$. In order to determine whether $\mathbf{x}$ or $\mathbf{y}$ is closer to $\mathbf{x}^*$ we can compare the ratio-vectors $\left(\frac{x_1}{x_1^*}, \ldots, \frac{x_{n^*}}{x_{n^*}^*}\right)$ and $\left(\frac{y_1}{x_1^*}, \ldots, \frac{y_{n^*}}{x_{n^*}^*}\right)$ (where $n_x^* = n_y^* = n^*$): in particular, we should choose the distribution corresponding to

the ratio-vector whose components are more "equal" or less "spread out". From majorization theory [14] we can identify the class of functions which are consistent with this principle by a weighted sum of increasing and concave functions of the ratios $\frac{x_i}{x_i^*}$.

In particular, we propose:

$$\Phi(\mathbf{x}) = \sum_{i=1}^{n_x^*} \phi\left(\frac{x_i}{x_i^*}\right) x_i^*, \tag{7}$$

where ϕ is increasing, concave but also positive and defined in 0. In the trivial case where a researcher has not received any citation (or published any paper), assume $\Phi(\mathbf{x}) = 0$.

It is of interest to note the relation between any function $\Phi(\mathbf{x})$ and the *relative majorization* (*r-majorization*) pre-order defined by Joe [22]. Suppose that $\mathbf{u}$ and $\mathbf{v}$ yield the same ideal distribution, say $\mathbf{q}$ ($\mathbf{u}^* = \mathbf{v}^* = \mathbf{q}$), and let $\widetilde{\mathbf{u}} = (u_1, \ldots, u_{n^*})$, $\widetilde{\mathbf{v}} = (v_1, \ldots, v_{n^*})$ (where $n_u^* = n_v^* = n^*$) so that $\widetilde{\mathbf{u}}$, $\widetilde{\mathbf{v}}$ and $\mathbf{q}$ have an equal number of elements n^*. Moreover, suppose that $\widetilde{\mathbf{u}}$, $\widetilde{\mathbf{v}}$ satisfy $\sum_{i=1}^{n^*} u_i = \sum_{i=1}^{n^*} v_i$ (equal citations within the n^*-core). In such a matching situation, the relation $\widetilde{\mathbf{u}} \prec_{\mathbf{q}}^{\mathbf{r}} \widetilde{\mathbf{v}}$, literally "$\widetilde{\mathbf{u}}$ is *r-majorized* by $\widetilde{\mathbf{v}}$ with respect to $\mathbf{q}$", means that $\widetilde{\mathbf{u}}$ is closer to $\mathbf{q}$ than $\widetilde{\mathbf{v}}$: thus $\mathbf{u}$ should be preferred to $\mathbf{v}$ (according to the basic logic set out in the previous subsection). It is proved that $\widetilde{\mathbf{u}} \prec_{\mathbf{q}}^{\mathbf{r}} \widetilde{\mathbf{v}}$ if and only if $\Phi(\mathbf{u}) \geq \Phi(\mathbf{v})$ ($\Phi(\widetilde{\mathbf{x}}) = \Phi(\mathbf{x})$ for any $\mathbf{x}$) for any concave function ϕ (note that this corresponds to the usual definition of *r*-majorization if we take $\phi = -\varphi$, where φ is convex). $\Phi(\mathbf{x})$ is said to be "order-preserving", "isotonic" [14, p.19] or *Schur-concave* with *r*-majorization [22], which means that if $\widetilde{\mathbf{u}} \prec_{\mathbf{q}}^{\mathbf{r}} \widetilde{\mathbf{v}}$ holds, then $\Phi(\mathbf{u}) \geq \Phi(\mathbf{v})$. In particular, ϕ is also non-decreasing because we cannot allow $\Phi(\mathbf{x})$ to decrease if an element of $\mathbf{x}$ increases (i.e. additional citations).

$\Phi(\mathbf{x})$ is based on the ratio between real/ideal citations per paper within the ideal set of citations i.e. the n^*-core. It is interesting to note the uncommon case when an author does not have enough publications i.e. $n_x < n_x^*$, which simply yields $\Phi(\mathbf{x}) = \sum_{i=1}^{n_x} \phi\left(\frac{x_i}{x_i^*}\right) x_i^*$ (the number of addends is inferior since we assumed that $x_j = 0$ for $j = n_x^*, \ldots, n_x$). Thus $\Phi(\mathbf{x})$ is indeed sensitive to the number of published papers. Moreover, the risk of considering papers which are not significant is countered by the fact that, if a paper has a low number of citations, the weight of those citations in Φ is downsized. On the other hand, $\Phi(\mathbf{x})$ is also sensitive to highly-cited papers, because ϕ is increasing. Nevertheless, for a fixed value of $C_x = C$, we obtain the best performance when $\mathbf{x}$ approaches $\mathbf{x}^*$, thus when the form of the distribution is "squared": this is consistent with respect to the basic logic of many bibliometric indices including the *h*-index (especially) and also the *g*-index (as proved above).

Within the general class defined by Φ, we choose $\phi(x) = \ln(1+x)$ (increasing, concave, positive and defined in 0), which yields:

$$L(\mathbf{x}) = \sum_{i=1}^{n_x^*} \ln\left(1 + \frac{x}{x_i^*}\right) x_i^*. \quad (8)$$

Finally, note that $h(\mathbf{x})$ and $g(\mathbf{x})$ are integer numbers defined on the interval $[0,\sqrt{C_x}]$. Thus, in order to obtain a bibliometric index which takes values within the same interval as the most popular ones (h and g), which can be useful for comparisons, we propose to normalize L as follows:

$$l(\mathbf{x}) = \sqrt{\frac{L(\mathbf{x})}{\ln 2}}. \quad (9)$$

Note that $0 \leq l(\mathbf{x}) \leq \sqrt{C_x}$, while h and g actually take values in $[0,\lceil\sqrt{C_x}\rceil]$.

$l(\mathbf{x})$ is based on a sum of a particular function that we denote by $\lambda(a,b) = \ln(1+\frac{a}{b})\ b$ ($a,b \in \mathbb{N}$ and $a,b>0$). In the S1 Appendix, we prove (*Lemma 1*) that $\lambda(a,b)$ is an increasing function of b (as well as a, obviously). This justifies and motivates the choice of $\phi(x) = \ln(1+x)$. The l-index outperforms the h-index in terms of precision and accuracy with respect to additional citations and sensitivity to the shape of the distribution. Moreover, l is (like h) robust with respect to citations in the set of non-significant papers. In particular, in the S1 Appendix the following properties are proved.

Property 1. Strict monotonicity with respect to citations

l is an increasing function of any additional citation.

Property 2. Robustness with respect to non-relevant citations

An additional citation within the n^*-core is always "heavier" than an additional citation outside the n^*-core.

Property 3. Sensitivity to regularity

An additional citation within the h^*-core is "heavier", the closer the cited paper is to the h^*-th paper.

Property 4. Sensitivity to elementary transfers

If $\mathbf{y}$ can be obtained from $\mathbf{x}$ by an elementary transfer of citations between two papers in the h^*-core, then $l(\mathbf{y}) \geq l(\mathbf{x})$.

Results

The main purpose of the paper is to find an index which improves the h-index in terms of its accuracy and sensitivity to both: i) citation "intensity" in the set of most significant papers; ii) the form of the citation distribution. For this reason, it is interesting to study the relations between l and some of the main alternatives to the Hirsch index (including the g-index and the h_T-index).

Theoretical examples

To verify the behavior of l we re-propose the theoretical examples provided by Vinkler [23], which illustrate the advantages and disadvantages of the h-index. The same particular cases were used by [8] to show the accuracy of the h_T-index. Before starting to analyze the results, we would point out that most of these theoretical datasets present quite uncommon features because they satisfy $h^*(\mathbf{x}) \geq n_x$. For this reason, in the next subsection we propose an application to real data.

The results in Table 1 show that the h_T-index improves the h-index (as already argued in [8]) by measuring both the quality and quantity of publications, but it is not very sensitive to highly-cited papers. For this reason, we also compute indices which are mainly aimed at assessing scientific impact such as the g-index, the p-index [24], the π-index and the R-index. Note that π and R are both based on the number of citations within a set of elite papers (respectively the π-core and the h-core). Moreover, we consider the z-index, an impact measure which is also sensitive to the form of the citation distribution and rewards regular (consistent) scientific performances.

On analyzing Table 1, first to be noted is that the g- and R-indices yield very similar results. More importantly, consider authors D and F: the h_T-score of author D is significantly higher than the h_T-score of author F. Conversely, the g-index is sensitive to the most cited papers but ignores the form of the citation distribution (authors A, D and F are equivalent according to their g-scores). Table 1 also shows that the π-index reflects scientific impact more accurately compared with the g-index in that it ranks author A above all the others and author F above author D. Indeed, on taking into consideration only the citations of the elite papers (i.e. the π-core), the π-index rewards a few papers of high impact in spite of poor regularity or consistency. Also note that, as mentioned above, $h^*(\mathbf{x}) \geq n_x$, so that every paper of every author (from A to F) belongs to the h^*-core; on the other hand, the number of elite papers considered for the computation of π is significantly smaller (e.g. 3 vs. 10 for author A), for this reason in this particular case the difference between l and π is especially accentuated.

The z-index behaves similarly to h_T and l if n_x is equal, this is because z is sensitive to the form (regularity). Conversely, when authors have similar numbers of citations but different numbers of published papers, a smaller number of papers may enhance the performance. In fact, the formula of the z-index is based on the product between a consistency measure and an impact measure, which is the p-index. In turn, the p-index is based on the ratio C_x^2 / n_x, where the number of papers is the denominator. Hence, among the considered indices, only p and z rank author F above the others, this is not just because of his/her number of citations but also because his/her number of papers is half that of the others.

The l-index seems to be "halfway" between the h_T -index and other impact measures because it is sensitive to both the form of the distribution and the number of citations of the most cited papers. Indeed consider again authors D

Table 1. Theoretical examples.

pap.\ aut.	A	B	C	D	E	F
1	100	9	10	50	9	10
2	98	8	10	50	8	110
3	98	8	10	50	7	100
4	97	6	10	50	6	90
5	96	5	10	50	5	80
6	4	4	10	50	_	_
7	3	4	10	50	_	_
8	2	3	10	50	_	_
9	1	2	10	50	_	_
10	1	1	10	50	_	_
n	10	10	10	10	5	5
C	500	50	100	500	35	500
h	5	5	10	10	5	5
h_T	13.27	6.89	10	18.5	5.79	12.46
g	22	6	10	22	5	22
R	22.11	6	10	22.36	5.91	22.36
π	2.96	0.25	0.3	1.5	0.17	2.3
p	29.24	6.29	10	29.24	6.25	36.84
z	23.54	5.82	10	29.24	6.17	36.59
l	17.03	6.7	10	19.54	5.76	16.64

Authors =A, B, F; *n*= number of papers; *C*= tot. number of citations.

doi:10.1371/journal.pone.0115962.t001

and F: according to l the gap between the scores of author D and author F is considerably reduced. On the other hand, l and h_T provide similar results when authors do not have highly-cited papers (authors B, E). The proposed l-index is strictly related to the h_T-index: l is sensitive to the "closeness" to the ideal distribution $\mathbf{x}^*$ which, as mentioned in the previous section, maximizes the h_T-scores. Nevertheless, there are some significant differences between l and h_T. Besides being sensitive to a "squared" form of the citation distribution, h_T is also *symmetric* (property defined by Kongo, [2]) while l does not fulfill the symmetry property (for Property 2 defined in the previous section). Indeed, to avoid any misunderstanding, we now prove that l and h_T are not *monotonically related* [25] with a straightforward counter-example. Consider $\mathbf{x}=(3,3,3,1)$ and $\mathbf{y}=(4,3,3,0)$: in this case $l(\mathbf{x})<l(\mathbf{y})$ but $h_T(\mathbf{x})=h_T(\mathbf{y})$. The l-index could be an improvement of h_T because it is sensitive to any additional citation and downsizes the effect of highly-cited papers (like h_T); on the other hand, it is not "symmetric" because the weight of the papers outside the n^* core (non-significant) is lower than the weight of the most cited ones (significant).

Case study

The Italian National Scientific Qualification (Abilitazione Scientifica Nazionale, ASN) is a new procedure, based on scientific qualification criteria, for the recruitment of academic staff in Italy. The ASN has involved tens of thousands of candidates (approximately 40,000). Here we focus on the set of 149 physicists who were applicants in the 2012 ASN for a full professorship in the specific area of Condensed Matter Physics. An expert panel of evaluators (a Committee of five members) was asked, by the Italian University Ministry, to approve (''habilitate'') or to reject each candidate. In Italy, habilitation is necessary to be eligible for a full professorship. The goal of the Committee was to select the best candidates by taking the impact of their scientific research into account.

The complete list of publications and corresponding citations for each of these applicants was retrieved by us from Scopus in January 2014. From the original (autoselected) sample of 149 datasets (for almost all the candidates for full professorship the status was that of ''Associate Professor''; the list of candidates was retrieved from the URL: http://abilitazione.miur.it/public/index.php), 18 datasets were discarded from the analyses due to insufficient citation data (e.g. an h-index less than 2) or difficulties in identifying the scientist. Then, for each of the 131 selected datasets, several different research productivity indices were computed, including l. We analyzed the results of h, g, h_T, l, but also π, R, p, z and the h'-index, recently proposed by Zhang as an index ''based on the citation distribution'' [10]. Moreover, we computed some simple bibliometric indicators such as the number of the citations of the most cited (MC) paper, the total number of citations C, the total number of papers n and the average number of citations per paper C/n. In Table 2 we present some descriptive statistics of the data. First to be noted is that, among 131 scientists, only 4 have a citation distribution such that $h^*(\mathbf{x}) \geq n_x$, confirming that this is a quite uncommon situation. However, for all the authors the total number of papers is always smaller than the number of citations, and also $h \geq \sqrt{n_x}$ except for only 2 of them. We therefore argue that, generally, the h^*-core includes the h-core, which in turn includes the π-core. Hence, in this situation the π-index is focused on the most elite papers (and therefore focused on impact), while the R-index, and consequently the l-index, considers larger sets of significant papers.

We also compared the results in terms of correlations between indices. Since in our opinion all those indices should be considered as measures at the level of *ordinal scale* and not *interval scale* (the critical question here is if the ''difference'' between, for example, two consecutive values of the h-index, h_0 and h_0 +1 scale, expresses the same ''gap'' regardless of the value of the baseline level h_0), these data should be analyzed only by using nonparametric methods for ordinal data. In particular, Table 3 presents the Spearman correlation coefficient (that is, the Pearson correlation coefficient between the ranked variables) for each pair of indices considered. As can be seen, the h'-index yields results which are not quite consistent with those of the other indices, in particular its correlation with the productivity index (n) is really low. More importantly, some indices show good

Table 2. Descriptive statistics.

pap.\ aut.	min	max	Mean	Q_1	Q_2	Q_3	SK	SD	CV
MC	5	3068	358	104.5	177	328	3.16	542	1.51
C	18	13916	2206	1156	1786	2716	2.49	1934.8	0.87
n	7	405	102	66	92	123	1.68	62.9	0.62
C/n	1.53	83.5	21.18	12.68	17.9	25.82	1.88	14.36	0.67
h	2	53	21.63	18	22	27	−0.10	8.66	0.40
h'	1.5	108.7	32.5	19.82	29.6	43	1.19	19.39	0.59
h_T	4.07	92.87	36.28	30.51	36.76	45	0.10	14.67	0.40
g	3	100	39.71	29	40	48.75	0.53	18.16	0.46
R	3.31	102.07	37.02	26.14	36.72	44.06	0.71	17.3	0.47
π	0.08	74.26	11.77	4.81	8.37	13.8	0.98	12	1.02
p	3.59	90.18	33.36	23.8	32.1	39.14	0.91	16.1	0.48
z	3.02	39.43	19.76	16	20.2	24.8	−0.14	7.36	0.3
l	4.03	98.73	37.51	30	37.82	46.3	0.26	15.56	0.41

Q_i =i-th quartile (i=1,2,3), SK = Skewness, SD = Standard Deviation, CV = Coefficient of Variation, MC = Maximum number of citations (x_1).

doi:10.1371/journal.pone.0115962.t002

correlation with C/n and therefore can be considered as impact measures: this set of indices consists of h', π, p, z, R and g (interestingly, g and R present very similar results, as already argued in [20]). In particular, some of these indices (h', π, p, R and g) are also highly correlated with MC, then, we argue that their values could be distorted by a single highly cited paper. On the other hand, h and h_T are also sensitive to the productivity, since they show good correlation with n. The l-index is highly correlated with both types of indices. Therefore, as hypothesized in the previous subsection, our data confirm that l is a good compromise for measuring both impact and form, indeed, it is especially appropriate for assessing authors based on the impact of their most cited papers as well as the regularity of their scientific production. To strengthen our thesis, it is also interesting to note that l is the index most correlated with C (l is a strictly increasing function of any additional citation, see property 1) and the second most highly correlated with n (after h_T).

Let us define the dichotomous "habilitation" variable, with values 0 (= rejected applicant) and 1 (= approved applicant). It is interesting to study the dependence between these indices and the judgements of the Committee (note that 69% of the 131 applicants were approved by the Committee). Table 4 reports the values of the Spearman correlation between the five indices considered and the habilitiation variable. Indices h, g, R and h_T show similar and good results in terms of coherence with the judgements; similar but slightly less satisfactory results are obtained for π, p and z; while the h'-index seem to be less associated with the habilitiation variable. Moreover, l is slightly more correlated with the habilitation variable than are h, g, R and h_T. Hence, we may suppose that l, which rewards reliability as well as the impact on the scientific community, reflects the evaluation criteria of the Committee in a quite satisfactory manner. Moreover, after

Table 3. Spearman correlation coefficients.

	h	g	h_T	h'	l	z	p	π	R	MC	C	n	C/n
h	1.000												
g	0.897	1.000											
h_T	0.971	0.920	1.000										
h'	0.674	0.858	0.669	1.000									
l	0.964	0.968	0.980	0.754	1.000								
z	0.871	0.863	0.853	0.769	0.872	1.000							
p	0.851	0.978	0.868	0.903	0.926	0.902	1.000						
π	0.834	0.982	0.876	0.860	0.931	0.786	0.955	1.000					
R	0.885	0.998	0.907	0.877	0.958	0.860	0.982	0.984	1.000				
MC	0.628	0.847	0.670	0.877	0.752	0.634	0.866	0.896	0.860	1.000			
C	0.927	0.985	0.955	0.779	0.986	0.850	0.947	0.966	0.979	0.802	1.000		
n	0.764	0.621	0.790	0.259	0.737	0.479	0.499	0.603	0.597	0.351	0.710	1.000	
C/n	0.624	0.803	0.622	0.944	0.702	0.819	0.892	0.786	0.820	0.834	0.728	0.153	1.000

Spearman correlation coefficients between bibliometric indicators.

doi:10.1371/journal.pone.0115962.t003

subdividing the sample into "approved" and "rejected" applicants, the W statistic for the two-sample Wilcoxon rank sum test [26] was also computed for each of the indices considered. We recall that the purpose of this test is to compare the ranks of one of the sub-samples (we considered that of the "approved" applicants: 91 cases) with those that would be expected if the null hypothesis of equal distribution of the levels of the index considered were true. The alternative is a condition of stochastic dominance, and, in our case, the null hypothesis was rejected for large values of W. Hence, one would expect higher values of W for the indices more in agreement with the Committee's judgement. Interestingly, as can be seen in Table 4, the Wilcoxon statistic W is strictly coherent with all the above results.

Conclusion

We have proposed a general method for improving the h-index that is based on the form of the citation distribution. The approach consists in defining an ideal

Table 4. Analysis of the habilitation variable.

	h'	z	p	π	h	g	R	h_T	l
HAB	0.450	0.533	0.569	0.576	0.590	0.S592	0.593	0.594	0.611
W	7033.0	7223	7304	7320	7349.0	7355.5	7361	7360.5	7400.0

First row: Spearman rank order correlation coefficients between the variable *HAB*" and various bibliometric indicators. Second row: Wilcoxon rank sum statistic (with reference to the cases in the larger of the two samples).

doi:10.1371/journal.pone.0115962.t004

optimal citation distribution for any author: a good bibliometric index should be sensitive to the closeness of the real citation distribution to its ideal one. In particular, the l-index is obtained when the reference distribution is ''squared''. Theoretical properties and empirical results from real data have been studied thoroughly. l rewards reliability and regularity, but it is also sensitive to highly-cited papers: its use is especially appropriate to evaluating (for instance) applicants for university positions, which is a major issue within the field of scientometrics. In particular, the statistical analyses on our case study yielded some interesting results: bibliometric rankings were compared with the judgments of a committee and it seems that l is the most appropriate (among the indices considered) for interpretation of this relation. Although the computation of l is not so simple (compared to the Hirsch index and some other popular bibliometric measures) the results of the paper are encouraging. They suggest that the new index could truly represent a significant alternative to the many existing h-type indices.

Supporting Information

S1 Appendix. Proofs.
doi:10.1371/journal.pone.0115962.s001 (PDF)

Author Contributions

Conceived and designed the experiments: LBB TL. Performed the experiments: TL. Analyzed the data: LBB. Wrote the paper: TL LBB.

References

1. **Hirsch J** (2005) An index to quantify an individuals scientific research output. Proceedings of the national Academy of Sciences of the United States of America 102 (46): 16569–16572.
2. **Kongo T** (2014) An alternative axiomatization of the Hirsch index. Journal of Informetrics 8: 252–258.
3. **Rousseau R** (2006) New developments related to the Hirsch index. Science Focus 1: 23–25.
4. **Zhang C** (2009) The e-index, complementing the h-index for excess citations. PLoS ONE 4 (5): e5429.
5. **Jin BH** (2006) H-index: An evaluation indicator proposed by scientist. Science Focus 1 (1): 8–9 (in chinese).
6. **Jin BH, Liang LM, Rousseau R, Egghe L** (2007) The R- and AR-indices: complementing the h-index. Chinese Science Bulletin 52 (6): 885–863.
7. **Egghe L** (2006) An improvement of the h-index: The g-index. ISSI Newsletter 2 (1): 8–9.
8. **Anderson TR, Hankin RKS, Killworth PD** (2008) Beyond the Durfee square: enhancing the h-index to score total publications output. Scientometrics 76 (3): 577–588.
9. **Prathap G** (2014) The Zynergy-Index and the Formula for the h-Index. Journal of the Association for Information Science and Technology 65 (2): 426–427.
10. **Zhang C** (2013) The h′-index, effectively improving the h-index based on the citations distribution. PLoS ONE 8 (4): e59912.

11. **Bertoli-Barsotti L** (2013) Improving a decomposition of the h-index. Journal of the American Society for Information Science and Technology 64 (7): 1522.
12. **Woeginger GJ** (2008) An axiomatic analysis of Egghes g-index. Journal of Informetrics 2: 364–368.
13. **Egghe L** (2010) The Hirsch index and related impact measures. ARIST 44 (1): 65–114.
14. **Marshall AW, Olkin I, Arnold B**, (2011) Inequalities: theory of majorization and its applications. Springer, New York, 2nd edition.
15. **Abbas AM** (2014) Bounds and inequalities relating h-index, g-index, e-index and generalized impact factor: an improvement over existing models. PLoS ONE 7 (4): e33699.
16. **Bornmann L** (2012) Redundancies in H Index Variants and the Proposal of the Number of Top-Cited Papers as an Attractive Indicator. Measurement 10: 149–153.
17. **Bornmann L, Marx W** (2013) How good is research really? Measuring the citation impact of publications with percentiles increases correct assessments and fair comparisons. EMBO reports 14 (3): 226–230.
18. **Vinkler P** (2009) The π-index: a new indicator for assessing scientific impact. Journal of Information Science 35 (5): 602–612.
19. **Vinkler P** (2010) The π_v-index: a new indicator to characterize the impact of journals. Scientometrics 82 (3): 461–475.
20. **De Visscher A** (2011) What Does the g-Index Really Measure? Journal of the Association for Information Science and Technology 62 (11): 2290–2293.
21. **Ali SM, Silvey SD** (1966) A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society 28 (1): 131–142.
22. **Joe H** (1990) Majorization and divergence. Journal of mathematical analysis and applications 148 (2): 287–305.
23. **Vinkler P** (2007) Eminence of scientists in the light of the h-index and other scientometric indicators. Journal of Information Science 33 (4): 481–491.
24. **Prathap G** (2010) The 100 most prolific economists using the p-index. Scientometrics 84: 167–172.
25. **van Eck NJ, Waltman L** (2008) Generalizing the h- and g-indices. Journal of Informetrics 2: 263–271.
26. **Wilcoxon F** (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1: 80–83.