Journal of Informetrics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Completing *h*

Keith R. Dienes

Department of Physics, University of Arizona, Tucson, AZ 85721, USA

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form
31 December 2014
Accepted 5 January 2015
Available online xxx

Keywords:
Citations
h-Index
Publications

ABSTRACT

Nearly a decade ago, the science community was introduced to the *h*-index, a proposed statistical measure of the collective impact of the publications of any individual researcher. Of course, any method of reducing a complex data set to a single number will necessarily have certain limitations and introduce certain biases. However, in this paper we point out that the definition of the *h*-index actually suffers from something far deeper: a hidden mathematical incompleteness intrinsic to its definition. In particular, we point out that one critical step within the definition of *h* has been missed until now, resulting in an index which only achieves its stated objectives under certain rather limited circumstances. For example, this incompleteness explains why the *h*-index ultimately has more utility in certain scientific subfields than others. In this paper, we expose the origin of this incompleteness and then also propose a method of completing the definition of *h* in a way which remains close to its original guiding principle. As a result, our "completed" *h* not only reduces to the usual *h* in cases where the *h*-index already achieves its objectives, but also extends the validity of the *h*-index into situations where it currently does not.

1. Introduction

In 2005, J.E. Hirsch introduced the so-called "*h*-index" as a way of assessing and quantifying the impact of the publication record associated with an individual researcher (Hirsch, 2005, 2007). Succinctly put, *h* is defined as the number of papers that the individual in question has produced which have at least *h* citations. Phrased less succinctly but perhaps more usefully, *h* is the maximum value of *N* for which it can be said that the individual has *N* papers with at least *N* citations each. The original motivation behind the definition of this index is that it balances between two opposite poles: the Scylla of total citation counts and the Charybdis of total numbers of papers. Although the quotient of these two numbers (the average number of citations per paper) is a useful measure for some purposes, it says nothing about how the citations are actually distributed amongst the papers – *i.e.*, whether they are all associated with just a few highly cited papers, or whether they are distributed fairly evenly across the publications, with no single publication attracting particularly strong attention. The *h*-index was therefore proposed as an alternative way of balancing between these two extremes and thereby assessing the overall "impact" of a given publication record.

It goes without saying that any statistical method of reducing a complex data set to a single number will necessarily have certain limitations that favor some researchers at the expense of others. Legitimate arguments can then be made for or against the proposed methodology, and in the case of the *h*-index a large literature devoted to this topic already exists.

It is not the purpose of this paper to engage in such discussions. Rather, in this paper we wish to point out that the definition of the *h*-index actually suffers from something far deeper: a hidden mathematical incompleteness intrinsic to its definition.

E-mail address: dienes@email.arizona.edu

http://dx.doi.org/10.1016/j.joi.2015.01.003
1751-1577/

In particular, we will demonstrate that one critical step within the definition of h has been missed until now, resulting in an index which only achieves its stated objectives under the rather limited circumstances in which the missing piece would not have had any effect. However, in other cases, it turns out that this missing piece is responsible for the apparent failure of h to act as originally desired. For example, we shall see that this incompleteness explains why the h-index apparently has more utility in certain scientific subfields than others.

Given this incompleteness in the definition of h, we then take the next step and propose a method of restoring the missing ingredient in a manner which remains consistent with the original guiding principles underlying h. As we shall see, this results in a new, "completed" version of the h-index, one which is mathematically robust across a wide variety of situations. Of course, our "completed" h reduces to the usual h in cases where the h-index already achieves its stated objectives. However, more importantly, our "completion" of h also extends its validity into situations where it currently does not.

2. Exposing the problem with *h*: a simple scaling argument

As described above, the h-index is designed to represent a rather ingenious *balancing* between paper counts and citation counts. Rather than focus exclusively on either total numbers of papers or total numbers of citations, h looks at how the set of citations is actually distributed across the set of papers, assessing the overall impact of a given publication record by seeking the point at which the number of well-cited papers matches the minimum number of citations those papers have. This balancing between paper counts and citation counts is the underlying motivation for h as well as the source of its ultimate utility. Unlike other proposed assessment variables, h is powerful because it represents neither variable exclusively but instead relies upon a subtle comparison of the two against each other.

However, it is easy to envision scenarios in which this balancing fails – *i.e.*, situations in which h ends up describing either a paper count *or* a citation count, with a value which is sensitive to only one of these variables and essentially insensitive to the other. For example, let us imagine two hypothetical scientists: one with 20 papers whose citation counts range from 1 to 20, and one with 20 papers whose citation counts range from 100 to 2000. In each case, the range of citation counts spans a factor of 20, and indeed the h-index of the first scientist is smaller than that of the second, as expected. However, we immediately see that the h-index of the second scientist reduces to a mere paper count, in the sense that further citations will have absolutely no effect on his h-index. By contrast, this will generally not be the case for the first scientist.

Although this example is trivial, it exposes the fact that the balancing inherent in h – indeed, its uniquely valuable feature – is vulnerable to situations in which paper counts and citation counts are of different orders of magnitude. In such cases, h entirely loses its sensitivity to one of these measures, and merely reflects the other. In such cases, the h-index has failed in its primary purpose, and no longer measures the subtle mixture of variables it was designed to assess.

Of course, the situation described above is somewhat contrived and unrealistic. Perhaps the most unrealistic aspect of the above example is the fact that every paper of our second hypothetical scientist has a citation count which exceeds his total number of papers. This is extremely rare, if it ever happens at all – in general, the citation counts achieved by a given scientist will range from some maximum value all the way down to zero. Indeed, implicit in the original definition of the h-index is the assumption that a given publication record will contain papers with number of citations both above and below h.

However, even under these more restrictive conditions, the overall scale associated with citation counts can still have the effect of destroying the balance inherent in h, thereby rendering h essentially insensitive to one variable or the other. To understand how this occurs, let us imagine ordering the publications of a given individual according to their citation rank r, so that the $r = 1$ paper has the most citations and papers with increasing r-indices have numbers of citations which either remain constant or decrease. Let us also assume that $N_c(r)$ represents the number of citations for each paper as a function of its rank r. In Fig. 1, we have illustrated the graphical means by which the corresponding h-index may be calculated: we simply calculate the point at which the $N_c(r)$ curve intersects the $r = N_c$ line. Fig. 1(a) illustrates the situation originally envisioned in Ref. Hirsch (2005), where indeed an almost identical figure appears: the overall scales for r and N_c are commensurate, so that the tangent line for the $N_c(r)$ curve is approximately perpendicular to the $r = N_c$ line. This implies that the resulting h-index represents a true balancing between numbers of papers and numbers of citations. In other words, the resulting h-index is just as sensitive to variations in the citation counts N_c as it is to variations in the paper rank r (as would occur if further well-cited papers were produced).

By contrast, in Fig. 1(b) and (c), we illustrate what occurs when the numbers of papers and the numbers of citations are of different overall magnitudes. Indeed, all we have done in passing from Fig. 1(a) to (b) and (c) is to rescale the overall $N_c(r)$ curve by an arbitrary small or large numerical factor. As evident from Fig. 1(b) and (c), this has the effect of rescaling the corresponding *slope* of the $N_c(r)$ curve at $r = h$ by the same factor. As a result, the tangent line for the $N_c(r)$ curve at $r = N_c(r) = h$ is no longer perpendicular to the $r = N_c$ line. Indeed, for particularly small or large rescalings of the $N_c(r)$ curve [as illustrated in Fig. 1(b) or (c), respectively], the slope of the tangent line at $r = N_c(r) = h$ tends toward either zero or (negative) infinity. In such cases, h becomes virtually insensitive to variations in either ranks or citation counts respectively.

This sensitivity issue is ultimately critical if h is to retain its original intended meaning. For example, if the $N_c(r)$ curve has nearly vanishing slope at $r = N_c = h$, as in Fig. 1(b), then the values of $N_c(r)$ with $r \gtrsim h$ will not be too different from $N_c(h)$. Consequently it will only take a sprinkling of relatively few additional citations to raise the corresponding h-index significantly. In other words, h will be extremely sensitive to small variations in citation counts. By contrast, if the $N_c(r)$ curve

Please cite this article in press as: Dienes, K.R. Completing *h*. *Journal of Informetrics* (2015), http://dx.doi.org/10.1016/j.joi.2015.01.003