Applied Network Science

RESEARCH Open Access

Fractal dimension analogous scale-invariant derivative of Hirsch's index

Yuji Fujita[1,2*] and Noritaka Usami[1,3]

*Correspondence:
yuji@turnstone.jp
[2] Turnstone Research Institute, Inc., Kamakura 2480004, Japan
Full list of author information is available at the end of the article

Abstract

We propose a scale-invariant derivative of the *h*-index as "*h*-dimension", which is analogous to the fractal dimension of the *h*-index for institutional performance analysis. The design of *h*-dimension comes from the self-similar characteristics of the citation structure. We applied this *h*-dimension to data of 134 Japanese national universities and research institutes, and found well-performing medium-sized research institutes, where we identified multiple organizations related to natural disasters. This result is reasonable considering that Japan is frequently hit by earthquakes, typhoons, volcanoes and other natural disasters. However, these characteristic institutes are screened by larger universities if we depend on the existing *h*-index. The scale-invariant property of the proposed method helps to understand the nature of academic activities, which must promote fair and objective evaluation of research activities to maximize intellectual, and eventually economic opportunity.

Keywords: Self-similarity, Complex network, Hirsch's index

Introduction

Citation structure of scientific literature is a major subject of the study of science policy and complex systems. Not only its graph-theoretic structure constitutes a valuable topic, but also its application has a significant impact on society.

The importance of citation structure comes from the fact that citing defines the flow of knowledge. The cited literature contributes to overcoming the intellectual limit of humans by supporting the following researches through the flow of knowledge. Therefore, studying the flow and diffusion of knowledge will allow us

1. to evaluate the contribution of scientific activity and,
2. to identify the knowledge flow to forecast the future of scientific research.

Moreover, because the flow of knowledge is a major principle of intellectual production, it also constitutes a prominent driving force of the "Knowledge-based economy" (see OECD 1996), where production, diffusion, distribution and transmission of knowledge is the key to the economic productivity.

In this study, we propose a derivative of J. E. Hirsch's index (Hirsch 2005) as ***h*-dimension**, which is analogous to the fractal dimension of Hirsch's index (***h*-index** hereafter) and has a property of being data-size invariant. *h*-dimension is developed as an institutional index, and it is *not* for individual researchers.

The authors, who are affiliated to Council for Science, Technology and Innovation (CSTI for short), a department of Japan Cabinet Office, are in charge of establishing evidence-based policy making (EBPM for short) by analyzing human and monetary resource investment, academic output and economic and social outcome. Such analysis is to be integrated to a data analysis platform system as "e-CSTI", which is now partially available to the general public as https://e-csti.go.jp/en. The purpose of this system is to analyze/evaluate governmental policy and act, and it is not to be used for judging or resource allocation of individual institutes or researchers.

In Japan there are more than 100 national universities and research institutes. These organizations have divergent research activities both in their scale and topics. Within the data set of this study there are approximately 1.0e+04 times of difference in the number of articles. The research topics are also highly diversified: major national universities' research fields are uniformly distributed from mathematics to human studies, but there are institutes specialized in a particular topic such as particle physics, medical science of a specific part of the human body, or genetics.

Among various methods and indices related to citation analysis, *h*-index proposed in Hirsch (2005) has an advantage in detecting stochastic features of citation data with robustness. Ever since, multiple variations of *h*-index are created to meet different objectives and data variations. Alonso et al. (2009) is a comprehensive review of such early works, where various applications of the original index and the derivatives are acquainted. Among more recent works based on the *h*-index is Amane Koizumi's *h5*-index (Koizumi 2018), which is gathering attention recently (see https://www.elsevier.com/__data/assets/pdf_file/0020/53327/ELSV-13013-Elsevier-Research-Metrics-Book-r12-WEB.pdf, page 40). *h5*-index was developed to measure institutional performance by setting a time-window of five years (hence the name "h5"). To be precise, let $X(t)$ be the set of articles produced by some organization from its founding up to some year t, then *h5*-index is defined as the *h*-index calculated on the article set $X(t) - X(t-5)$ provided that $t \geq 5$. The definition of *h*-index is given in "Self-similarity of citation and h-dimension" section. *h-5*index is always associated with a particular interval of five years, for example, from 2014 to 2018, and it can be any five years if adequate data is available. Koizumi's work also directly inspired this study.

Still, we see difficulties in using *h*-index (or *h5*-index) for institutional performance evaluation because these indices are heavily correlated to the number of articles on which the indices are defined. It means we cannot compare institutes with a different number of researchers. Another difficulty comes from the fact that the network of citation acquires vertices and links along with time, and each researchers organization has their own stage of development.

Self-similarity or fractal-like property is almost a universal concept to analyze and understand complex structures. We consider this concept as a primary principle to analyze citation networks. We developed our version of *h*-index derivative *h*-dimension, or

h_d to take advantage of this property. As a result the index is necessarily scale-invariant to bring fair and accurate institutional evaluation.

There is substantial criticism on the evaluation of individual researcher based on h-index. In Koltun and Hafner (2021), declining effectiveness of the scientometric measures is reported, that correlation of h-index with scientific awards has dropped due to the changing authorship patterns. In Waltman (2016), it is claimed that productivity over expenditure, or achievement per budget is the key factor to be evaluated, and all the scientometric index which does not consider monetary cost is meaningless or harmful. In Waltman and van Eck (2012), it is claimed that h-index is prone to noise, which makes the index to lose the order-preserving property and consistency.

However, because this study is strictly institution-oriented, we consider such criticism is not very relevant. Moreover, we believe that a set of academic articles which are richly linked by citation is more valuable than disconnected one, and creating such knowledge circulating system is an important mission of research institutions.

The present study is constructed as follows: in "Self-similarity of citation and h-dimension" section, we will examine the structural characteristics, in particular statistical self-similarity of citation. This examination leads us to the fractal dimension of h-index as h-dimension, or h_d. In "h-dimension and its implication" section we will apply the proposed index h_d to the prepared data to study the properties and implications of the proposed index. By referring to h_d, we found well-performing medium-sized institutes, which are obscured by larger organizations if we depend on the original h-index. We also check the effect of the research field on h_d, to find out that it is not heavily affected by the research field selection. In "Conclusion and future work" section we will summarize this study and discuss the future work.

Self-similarity of citation and *h*-dimension

In this section, we propose "h-dimension", a scale-invariant derivative of the h-index, by considering the self-similar characteristics of the citation network. It is analogous to the so-called fractal dimension of the h-index, hence it is named as such.

As we briefly discussed in "Introduction" section, the citation network represents the propagation of knowledge, which means understanding structural characteristics of the citation network will give us insight into the flow of knowledge. For the purpose of structural discussion, we begin with formulating the citation network.

Let $X = \{u, v, \ldots\}$ be a set of articles we are concerned, and $E = \{(u, v), (u, w), \ldots\}$ be a set of citation relations between the elements of X; i.e., $(u, v) \in E$ implies $(v, u) \notin E$ because citation is asymmetric. Let us note that a citation relation (u, v) means the article u is cited by the article v. The pair of sets (X, E) defines the "citation network" as an acyclic directed graph. Let c be a function from X to Z,the set of natural numbers, as $c(u) = |\{v|(u, v) \in E\}|$; function c counts how many times u is cited, or in-degree of the vertex u.

Let D be a distribution function in general. For simplicity, let $D(b)$ be the possibility to be $x < b$ provided that x is a probabilistic variable on which D is defined.

h-index h is defined as follows; let $U(x)$ be a subset of X such that all the members of $U(x)$ has x or more incoming links, then

$$h = sup\{x|\ x \leq |U(x)|\}. \tag{1}$$

Equation 1 can be written more simply by using empirical distribution function D as follows; let $n = |X|$, the number of articles, and $\mathbb{D}$ be a set-valued function as $\mathbb{D}(x) = \{u|\ u \leq nD(x)\}$. Then h can be written as

$$h = sup\ \mathbb{D}(h). \tag{2}$$

Note that Eq. 2 describes the h-index as a fixed point based on an empirical distribution function, and this fact directly yields an effective h-index calculation algorithm deployed in this study, which is described in "Appendix" with a sample code.

Let H be the subset of X which defines h-index, namely,

$$H = \{u|c(u) \geq h\}. \tag{3}$$

The citation network evolves with time by adding a new article to the network. Therefore, it is natural to identify the increment of n, the size of the data, or $|E|$ with ongoing time t, which is a common approach to model how a citation network (or other real-life complex network) is built (see Price 1976 or Barabasi and Albert 1999 for example).

Each research institute or university has its own history. An older institute is likely to have a larger set of articles if the researchers' number is similar. On the other hand, the expectation of the h-index of a larger set of articles is larger if the citation distribution is the same[1]

The median of personal h-index of the researchers who are affiliated to the institute is a good candidate for the purpose of institutional performance measurement. However, the median is not available because there is a significant number of research-related people who never appear as the author of the academic papers. The proposed index of this study has the advantage that it only depends on the academic paper database.

Therefore, a simple comparison of h-indices of different institutes' data may only mean that one data has more articles, which means raw h-index is inadequate for institutional evaluation. Actually, the same difficulty exists in the case of personal h-index, which is addressed in the original paper of Hirsch (2005) by denominating the raw index value by the years of being an active researcher.

$h5$-index, which sets a fixed length of five years window to collect the data, overcomes this difficulty by taking a snapshot of uniformly controlled exposure. Despite this improvement, we still have the following concern about the growth of the citation network.

A research institute goes through its own process of development. Newly established institute N is in its early stage, while another institute M is in its mature stage. Suppose if these two institutes share a similar index value. If we are to conclude that N and M achieved similar performance based on this index, the conclusion has very limited significance because N is doing better. Comparison between cases with similar stages of development can be meaningful; otherwise, we are uncertain despite the controlled observation window.

[1] This can easily be shown from Eq. 2 and the fact that distribution function D is monotonically decreasing. See (PRATELLI et al. 2012) for detailed discussion.

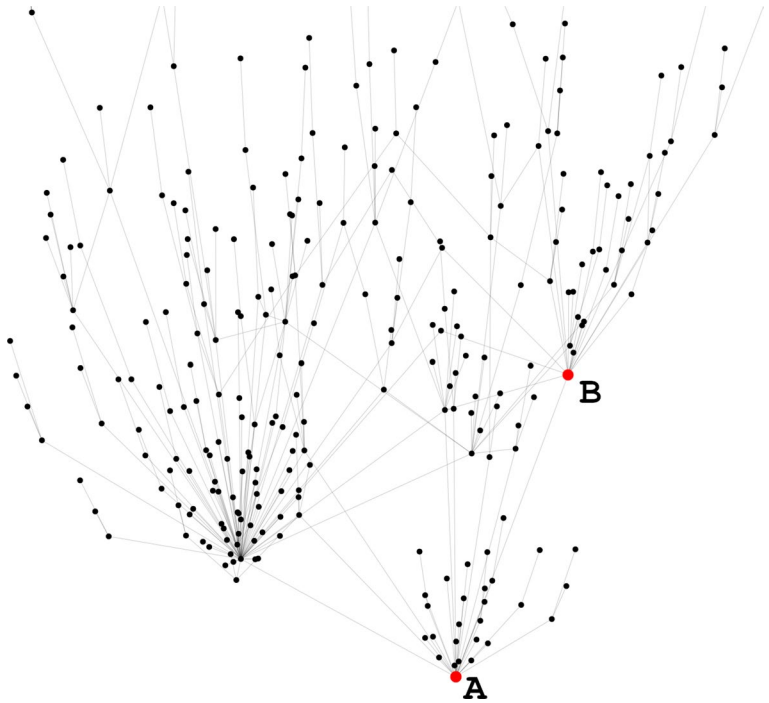

Fig. 1 A citation network artificially constructed for the purpose of illustration. Each dot represents an article, and the straight line is a citation relation. To visualize citation direction, a cited article is placed below the citing article. The article "A" (bottom red dot) comes first within this network, then "B" comes after it to cite "A". The network above "A" (the whole network) is similar to the one above "B" in a sense that they share some statistical properties like degree distribution

Figure 1 is a visualization of an artificially created acyclic directed graph with similar in-degree distribution of typical citation network. The layout is configured to place cited article below its citing article. Because the citing article necessarily comes after the cited one, the network "grows" upwards like a tree grows to the sky.

Inside Fig. 1, we can identify a sub-network by selecting all the nodes that refer to "B" directly or indirectly. This sub-network is *similar* to the original network (which is the articles that cites "A" directly or indirectly) in the sense that they share statistical properties, for example, the degree distribution. Benoit Mandelbrot refers to this property as *statistical self-similarity* in his book (Mandelbrot 1977), Chapter XII, p. 276.

In fact, the fractal-like property of the complex network has been gathering attention (for example Song et al. 2005; Corominas-Murtra et al. 2013 or Zhao et al. 2006). Hierarchical characteristics of acyclic network is sometimes called "rank" structure from the fact that it is often embedded in a one-dimensional ordered space (see Newman 2018 14.7, p. 564).

The h-index defining set H of Eq. 3 and their associated links E_H also defines a sub-graph (H, E_H), which occupies the lower part of Fig. 1 . The in-degree distribution D_H of this sub-graph is obtained from the original distribution function D as $D_H(u) = \frac{D(u)}{D(h)}$ provided that h is the h-index value of the whole citation network and $u \geq h$. Therefore, the sub-graph (H, E_H) shares degree distribution with the original network (X, E) except that it lacks the long-tail, or lower degree part of the distribution.

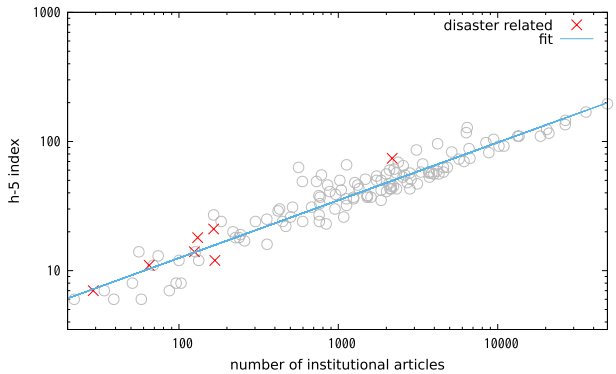

Fig. 2 Japanese institute-wise *h5*-index v.s. the number of articles. The fit function in the plot is $y = 1.6x^{0.47}$. *h5*-index is correlated to the number of articles by coefficient of +0.85. The disaster-related research institutes are plotted in red "x"

Obviously, the empirical data necessarily have finite steps of similarity, i.e. we cannot go down to statistically similar sub-networks infinitely. In other words, the level of detail of the observed network is limited. Due to this self-similarity and nature of citation relation, adding a new node (or growth of the network) also means adding detail to the network.

If the sub-network H converges to a stationary state when $t \to \infty$, such a terminal state should consist of a fair evaluation foundation indifferent to data scale or the institute's history. However, it is unlikely that such a stationary state exists because of the statistical self-similarity of the citation structure.[2] It means infinitely detailed observation leads to infinitely larger observed value, which makes comparison impossible. This difficulty is analogous to the measurement of coast line length, which famously diverges to infinity as the mesh of the survey becomes smaller for more detailed measurement. (see Mandelbrot 1977, chapter 2).

Fortunately, we already know how to treat a measurement of self-similar structure that diverges with the observation scale, which is the *fractal dimension*.

According to this knowledge, we propose h-dimension h_d as follows:

$$h_d = \frac{\log(h)}{\log(s)}, \tag{4}$$

where h is the h-index (or $h5$-index) value, which is the objective measurement, and s is the size of the network, which is the inverse of the scale of observation. The number of articles can not be used as s because it only counts the vertices which have incoming links. In practice, s can be obtained as the sum of the citation counts $\sum_{u \in X} c(u)$.

[2] Statistically similar sub-network keeps to grow along with the whole network unless the similarity is broken at some point. In fact, in Fig. 2 of "h-dimension and its implication" section, we will see that the stationary terminal state existence is empirically denied.

h-dimension and its implication

In this section, we examine the properties and implications of h_d. In the first subsection, we describe the preparation process of the data. In "Application and examination of h-dimension" section we apply the proposed index h_d to the collected data and analyze the theoretical and empirical properties of h_d. We will check if it is scale-invariant. In "Research field and h-dimension" section we will examine the effect of the research field selection on h_d, and give some intuitive understanding of the proposed index through statistical analysis. In "Adversary strategy against h-dimension" section, we will examine h_d from the "opposite side of the game" and try to construct the strategies to deceive the index.

Data preparation

We used a bibliometric dataset for research articles published in 2014-2018 from national university corporations, national research and developments agencies and inter-university research institute corporations in Japan. As a starting point for data collection, a list of institutional identifiers was prepared by using a global research identifier database (Data-Science I 2019). As a consequence, 134 GRID (Global Research Identifier Database) ids were obtained including all 86 national university corporations in Japan. We used the Dimensions analytics API (https://docs.dimensions.ai/dsl/) as the platform, which permits us to extract fundamental bibliometric data on a research institute for a given period related to a specific research field by a simple query. The query language is not very different from a common SQL with some extensions.

Since we could extract up to a maximum of 1,000 results only, it is necessary to add the operator, skip, followed by the offset, if we like to obtain all the data when the total count of the result is over 1,000. This iteration to give the offset could be done up to a maximum of 50,000 results. No queries for a specific research institute and the research field gave results of more than 50,000. Therefore, by iterating such queries for 134 research institutes and 22 research fields, we successfully collected the dataset for further analysis.

The number of citations for each article is as of 14 March 2020 in Dimensions, when we collected all the data. The dataset of this study consists of citation counts of 550,602 papers. To measure *h5*-index for each research institute, a set of unique publication ids over all research fields and their number of citations was used. Then *h5*-index was measured according to the definition given by Hirsch (2005), using the algorithm addressed in "Appendix". Technically, the only difference between the *h5*-index and *h*-index is that we used research articles for 5 years from 2014 to 2018 in order to measure the recent activities for a research institute instead of the whole activities.

Application and examination of *h*-dimension

Figure 2 shows 134 institutes' *h*-index (*h5*-index) values and their number of institutional articles in a log-log plot. It is clearly seen that the *h5*-index and the number of articles shows a strong positive correlation, whose coefficient is 0.85. Figure 2 also empirically denies the existence of the stationary terminal state of the *h*-index defining sub-network *H*, which is discussed at the end of "Self-similarity of citation and h-dimension" section.

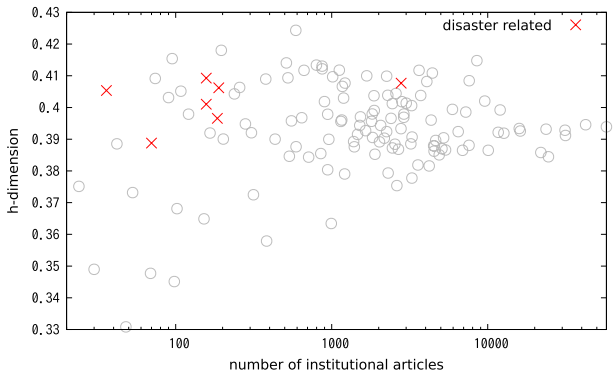

Fig. 3 Japanese institute-wise *h*-dimension v.s. the number of articles. The index is correlated to the number of articles by +0.07. The disaster-related research institutes are plotted in red "x"

Figure 3 shows h_d versus the number of articles plot, where we can see that the scale-dependence problem of the *h*-index is mostly resolved. The result has a smaller correlation coefficient of +0.07 with the number of articles.

Thanks to this improvement, we could find well-performing medium-sized organizations with the number of institutional articles from several hundred to ten thousand. Among these organizations we could identify several research institutes focused on natural disasters, as shown in red "x" in Figs. 2 and 3. There are several institutions that marked even better than disaster-related organizations, to which we do not refer any further to avoid making them identified.

This result is reasonably understood because Japan is frequently hit by various natural disasters, like typhoons and earthquakes. In contrast, these characteristic leading institutes are overshadowed by larger organizations in Fig. 2. It is evidently seen that the proposed index is useful to evaluate institutional performance by a relatively simple calculation.

However, we cannot easily conclude that the result of Fig. 3 is actually representing *institutional* performance. The organizations in the data have different research fields, which are known to have different citation conventions.

In the following part, we will show that the proposed index h_d is robust to the research field configuration difference. This is due to the scale-invariant design of h_d and property of the original *h*-index to select an essential part of the data.

Research field and *h*-dimension

In general, research field is a major factor of the outcome of citation, for example in Qian et al. (2017) it is shown that even the sub-fields within computer science have significant effect on the citation rates. Various field-normalized bibliometric methods are devised and utilized to compensate such variation for the purpose of fair and accurate evaluation (see Ahlgren and Sjögårde 2015; Bornmann and Haunschild 2016; Reddy et al. 2020).

To begin with, we estimate the effect of the research field selection on h_d. The value of *h*-index is defined by the distribution of citation as described in Eq. 2. The data we are analyzing actually has a joint distribution of the research institutes, research fields, and citation count. Here we define three probabilistic variables as follows:

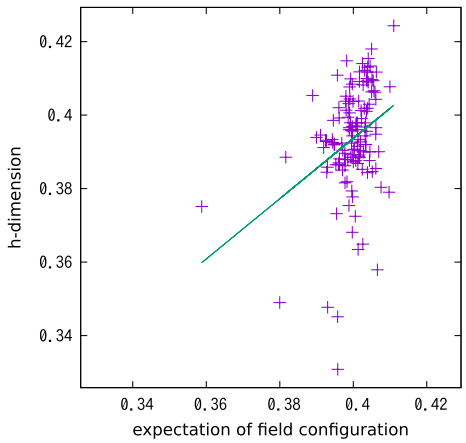

Fig. 4 h_d values of controlled expectations of research field configuration v.s. the data. The fit function of green segment is $y = 0.8x + 0.07$. Kendall's rank correlation is 0.2 and $\frac{Residual}{Data\ Variance}$ is 0.88

let the research fields be F, institutes G, and citation C. Let $P(X)$ be the probability that a statement X is true, namely, $P(F = f, G = g, C = c)$ denotes probability of $F = f, G = g$, and $C = c$ for some research field f, institute g, and citation value c.

Then the distribution function $D(h)$ of Eq. 2 can be written as the marginalization of the original data: $D(h) = \sum_G \sum_F P(F, G, C < h)$. The research field distribution of some institute g_i can be written as $\sum_C P(F, G = g_i, C)$. To see the effect of research field difference, we must obtain citation distribution which is independent of particular institute g_i. This can be performed by calculating

$$\sum_F P(F, G = g_i) P(C < c | F = f) \tag{5}$$

of the data. Aggregated term of Eq. 5 is a conditional distribution of research field f multiplied by the proportion of the field f of the research institute g_i. Note that Eq. 5 uses distribution of all the articles that belong to research field f, and not the observed distribution of particular institute g_i. Adding this term over research fields will give randomly controlled citation distribution under the condition of research field selection of particular institute g_i.

Figure 4 is a scatter plot of h_d of randomly controlled data of Eq. 5 (the horizontal axis) and corresponding institute's observed h_d value. In practice, the control is obtained by averaging 200 runs of the randomly selected results. The plot shows weak positive correlation between the controlled expectation and observed value, which is represented by the green segment. However, the residual reached nearly 90 percent of the observed variance. We also checked Kendall's rank correlation coefficient between the controlled expectation and the data, which was 0.2. In summary, only 10 to 20 percent of the institutional h_d variation is explained by research field selection. Much of the h_d variation comes from outside of the research field selection.

We will further analyze how this almost research-field invariant property of h_d is realized. Table 1 is a list of research-field wise values of the number of articles, mean

Table 1 Research field-wise number of articles, sample mean of citations, *h5*-index and *h*-dimension from the prepared data set

Research field	Number of articles	Average citations	*h5*	h_d
Studies in the creative arts and writing	207	3.52	13	0.38
Philosophy and religious studies	490	1.77	13	0.37
Law and legal studies	704	2.76	18	0.38
History and archaeology	1128	3.64	36	0.43
Language, communication and culture	1327	2.28	22	0.38
Built environment and design	1482	5.58	36	0.39
Education	1557	2.61	23	0.37
Commerce, management, tourism	1740	5.89	38	0.39
Economics	3206	6.18	51	0.39
Studies in human society	3373	4.6	48	0.4
Environmental sciences	6624	10.87	84	0.39
Agricultural and veterinary sciences	9025	6.07	61	0.37
Technology	10,648	9.83	93	0.39
Psychology and cognitive sciences	10,975	5.92	71	0.38
Information and computing sciences	15,387	5.97	87	0.39
Mathematical sciences	18,286	6.76	98	0.39
Earth sciences	19,089	11.06	120	0.39
Physical sciences	64,809	12.74	210	0.39
Chemical sciences	66,265	12.38	203	0.39
Biological Sciences	70,363	14.73	253	0.39
Engineering	88,527	8.83	182	0.38
Medical and health sciences	155,390	10.31	255	0.38

citation, *h5*-index, and *h*-dimension from the prepared data as described in "Data preparation" section.

We can see that mean citations in the 3rd ("Mean") column and *h5*-index in the 4th column have great variation. Also, these two indices have a strong positive correlation with the size of the data.

In contrast, the proposed index h_d listed in the far right column of Table 1 shows nearly constant values.

As we discussed in "Self-similarity of citation and h-dimension" section, the original *h*-index expectation is a monotonically increasing function of the number of articles. Therefore, the positive correlation of *h*-index to the number of articles in Table 1 is natural.

However, as for h_d, which is scale-invariant by its design, is not guaranteed to yield uniform values because different research field may have different citation distribution.

In Fig. 5, we can see that each research field shows approximate power-law distribution above its *h*-index value[3]. Although these distributions have difference in their fewer times cited part, the *h*-index defining part have very similar distributions with each other. We will see this relation in more detail with simple analytic calculation and empirical check as follows:

[3] *h*-index is at the crossing-point of the diagonal straight line and cumulative distribution curve.

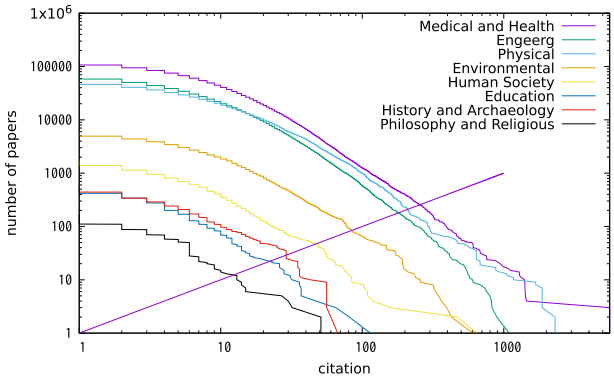

Fig. 5 Research field-wise citation distribution plots. Note that vertical axis shows cumulative article count, not cumulative proportion for the purpose of easy distinction. The diagonal straight line from upper right to lower left corner is an aid to find their corresponding *h5*-indices. Above the *h5*-index value, each research field shows similar power-law distributions (except for History and Archaeology). Power-law exponent values are just above 1.0

The h_d definition of Eq. 4 can be transformed to $h = s^{h_d}$, where h_d serves as the exponent over the inverse of the observation scale s. Then, $s = n^{\frac{a}{a-1}}$ because the expectation of power-law distribution with exponent a is $\frac{a}{a-1}$. This is not valid in the overall distribution of citation data, but it holds above the h-index defining point.

Consequently it holds that

$$h = kn^{h_d} \tag{6}$$

for a constant $k = \frac{a}{a-1}^{h_d}$.

By substituting $D(x)$ of Eq. 2 with bx^{-a}, where b is a constant for normalization to be compatible with the probability distribution, we obtain $h = nbh^{-a}$, which is transformed to

$$h \sim n^{\frac{1}{a+1}}. \tag{7}$$

Equation 7 is not a new result, and already described in PRATELLI et al. (2012). From Eqs. 6 and 7 we have $n^{\frac{1}{a+1}} \sim kn^{h_d}$. Therefore

$$h_d \sim \frac{1}{a+1} \tag{8}$$

is obtained.

To confirm the relation between h_d and power-law exponent of Eq. 8, we estimated the power-law exponent value of 134 data sets by applying Hill's estimator (Hill 1975) and compare them with h_d.

The estimation result is shown in Fig. 6, where we can see its mode is approximately $a = 1.2$. Alternatively, we estimate the power-law exponent by way of the h-index. Compare Eq. 7 with the fit function of Fig. 2 as $0.47 = \frac{1}{a+1}$, which gives $a = 1.14$. Two different estimates meet quite well.

However, it should be noted that the exponent of the power-law distribution and h_d are based on totally different principles. The power-law exponent is a result of *fitting* a

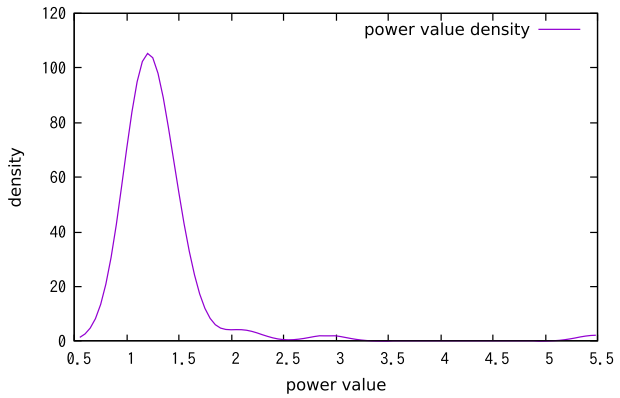

Fig. 6 Estimated power values of the Japanese institutes' data set distribution. The plot shows probability density to see the mode, which is approximately 1.2

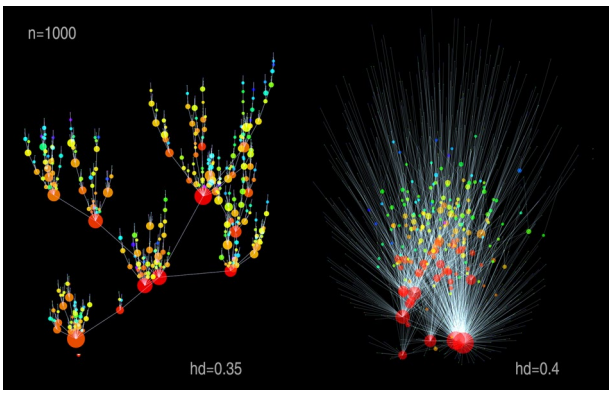

Fig. 7 Artificially constructed two acyclic directed graphs with the same number of vertices and different h_d values. The graphs are built by adding a vertex repeatedly, just like the real-life complex network. The older vertex in red, then yellow, green, and the new one in blue, and the vertex size is proportional to its in-degree. The other layout configurations are identical to Fig. 1

particular function to the observed data. h_d, on the other hand, is based on nonparametric statistics of h-index and fractal.

In summary, h_d of the research fields are stable because all the research fields share very similar distributions above the h-index defining values, which we can see in Fig. 5.

Following the definition of h_d and self-similar characteristics of citation network, we can infer that a network generated by unifying multiple statistically similar networks (which means these two networks share the same h_d values) will also yield the same h_d value. This is consistent with the fact that the original networks are statistically similar subgraphs of the newly generated network. In other words, h_d is invariant to the operation of identically distributed data set unification.

As a result, we can claim that the difference in research field selection has limited effect on the institutional h_d, which means most of the h_d variation is not originated from their research fields.

If we stay in the realm of self-similar citation structure, Eq. 8 will give us a rough idea of what h_d is trying to measure. High h_d, or small exponent of power-law distribution

implies that there is an active and effective knowledge production process which provides readily available knowledge with quality and quantity to create new knowledge. In terms of citation structure, it is a richly connected network.

A pair of artificially constructed acyclic directed graphs with the same number of vertices and different h_d values are visualized inFig. 7. The left (lower h_d valued) graph consists of several *communities* which are sparsely connected with each other, and the higher h_d graph is more tightly connected. If these two graphs represent citation relations, the higher h_d graph seems to have more foundational research activities shared among the researches than the lower h_d valued one.

Adversary strategy against *h*-dimension

In this subsection we will discuss the property of h_d by trying to conceive a work-around path to gain the index.

Following the Eq. 4, h-dimension is defined as

$$\frac{\log(h\text{-index})}{\log(\sum citations)}. \tag{9}$$

Therefore, the basic heuristics of the strategy is

1. to gain the h-index value, and
2. to keep the sum of citations small.

For a given citation sum C, the largest possible h-index is the integer part of $\sqrt{C}$. Thus, h-dimension has a range of $0 \leq h_d \leq 0.5$.

The network with its maximum h_d value has a unique degree distribution. Suppose the h-index defining set has h vertices, then all of these vertices have in-degree of h, and no other vertex has in-degree. Let us refer to the network that has h-dimension of 0.5 as "h_d -optimized".

In a naturally grown citation structure, when an article is cited it becomes more likely to be found and cited by other researchers. But in order to reach the h_d-optimized status, natural citation mechanism have to be stopped in order to avoid citation beyond the targeted h-index value, because any citation that does not contribute to the h-index brings down h-dimension.

Another way to gain h_d is to decrease the denominator while keeping the nominator of Eq. 9 by eliminating articles which does not contribute institutional h-index. This is unlikely to be implemented because all the newly published articles have no citations.

These two by-pass strategies are naive, and much more sophisticated work-around will be devised eventually. But for the time being, we consider that h_d is not particularly easy to work-around if we stay in the world of self-similarity.

Conclusion and future work

We proposed an index h-dimension, or h_d as a derivative of the h-index which is analogous to the fractal dimension of the original h-index. Unlike the original h-index, not only h_d is invariant to the number of articles by its design, it is also robust to the difference of research fields, if not completely independent.

Due to the self-similar property of the citation network structure, h-index is strongly and positively correlated to the number of articles, which gains its size as the time goes on. Most of the difficulties in comparing research organizations comes from this fractal property, and we already have an excellent tool to analyze this problem as "fractal dimension", which was named as such by the famous researcher Benoit Mandelbrot. h_d is defined as the fractal dimension of h-index.

We prepared a citation data set of 134 Japanese national universities, national research and developments agencies and inter-university research institutes from the year 2014 to 2018 by using Dimensions analytics API of DigitalScience, Inc., and applied h_d to the data. We could find several medium-sized research institutes that performs excellently by virtue of scale-invariant property of h_d, where we could identify multiple organizations focused on natural disaster, which is a reasonable result considering the natural environment of Japan. These characteristic institutes are obscured by major organizations if we depend on the conventional h-index.

We carried out several analysis from various angles on the properties of h_d. We examined the effect of research field on h_d value, to find out that the original h-index (therefore h_d as well) is closely related to the exponent of power-law distribution, which is quite similar among research fields above their h-index values. This is the reason why h_d is quite robust against difference in the research field. This property can practically exclude the effect of research field difference from institutional h_d values.

We also examined how visually different a graph with higher h_d is from a lower one, to find out that lower h_d graph is separated into multiple loosely inter-connected communities. In order to understand the behavior of h_d from a different point of view, we also tried to "attack" h_d and gain the value without following the supposed citation structure growth procedure, to find out that the mechanism behind self-similarity of citation structure is natural and hard to destroy.

International comparison of research organizations based on h_d must be an interesting research topic. The contraposition of the discovery of medium-sized institutes also constitutes a future investigation theme, i.e., the reason why the organizations which produced more than 1.0e+04 articles achieved consistently low h_d values.

Although single measure can never be the final solution to achieve fair and accurate institutional evaluation, we believe the h-dimension can help making good strategic decision to maximize intellectual or economic opportunity. We hope that the network study based on the principle of fractal to become even more active, and if this study could encourage it, the authors could not be more delighted.

Appendix

Here we describe our effective h-index calculation algorithm which is used throughout this study for quick and repeated analysis trials. It is expected to work in $O(n)$ of time to the input size of n.

Typical straight process to obtain h-index from observed data set X is as follows: sort the data X in descending order as $(x_0, \ldots, x_n)$. Then find the first element whose value is less than or equal with its rank, that is, find the first x such that $x_h \leq h$. Here

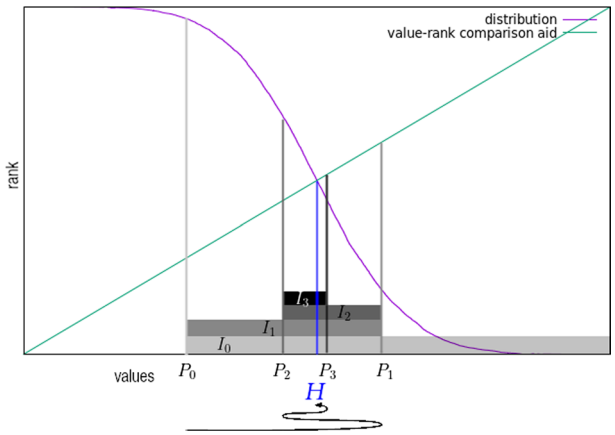

Fig. 8 An intuitive outline of the algorithm. The vertical axis is for the rank, up for larger, or lower rank. The horizontal axis shows the value of the data. Purple curve shows the empirical distribution function of the data. For the aid of comparison between the value and its rank, green diagonal straight line is added. To calculate the h-index value is to find out the point where the rank and value is equal (or nearest), which is the crossing point H. In the beginning, P_0 is randomly selected from the whole data. Hence P_0 has larger rank than the value itself, H must be in the segment I_0, $(\geq P_0)$. Next value P_1 is randomly selected from I_0. P_1 has a smaller rank than its value, therefore H should be in the segment I_1, $(\leq P_1)$. Then, P_2, which has larger rank than its value is randomly selected from I_1. P_2 sets the next segment I_2. Repeat the process two more times to finally reach H. Black slithering curve at the bottom of the figure depicts the search path

the rank h of such x_h is the value to be obtained. We will refer to this process as "sorting method" hereafter.

The major part of computational cost of sorting method exists in sorting the data set X, which is expected to be $O(n\log(n)))$.

Advantage of the sorting method is that it is reasonably fast and straightforward. The problem is the ineffectiveness. Sorting of h-th element beyond is not necessary. Actually, first h elements have no need to be sorted either.

The algorithm described here is a direct consequence from the fact that h-index is a fixed point of empirical distribution function, which is formulated as Eq. 2 of "Self-similarity of citation and h-dimension" section.

Figure 8 illustrates the outline of how the proposed algorithm works. The value to be calculated rests on the crossing point of diagonal green line and the empirical distribution function (purple). To find out the point, we start from a randomly selected value P_0 from the whole data, and repeat the following process: Randomly select P_i from a data segment I_{i-1}. If P_i has larger (or smaller) rank than P_i itself, the fixed point is in the segment above (or below) P_i, which is set as I_i and repeat this process to eventually reach the fixed point of Eq. 2.

However, from Fig. 8, it may seem that the algorithm requires totally sorted data. Therefore we will give a non-visual description in the following part of this section.

In the context of information science, proposed algorithm is a derivative of an algorithm commonly referred to as "quickselect", which outputs $k-$th largest (or smallest) element from given data of size n in average computational time of $O(n)$. Quickselect itself is a variation of quicksort, both were developed by the same person C. Hoare

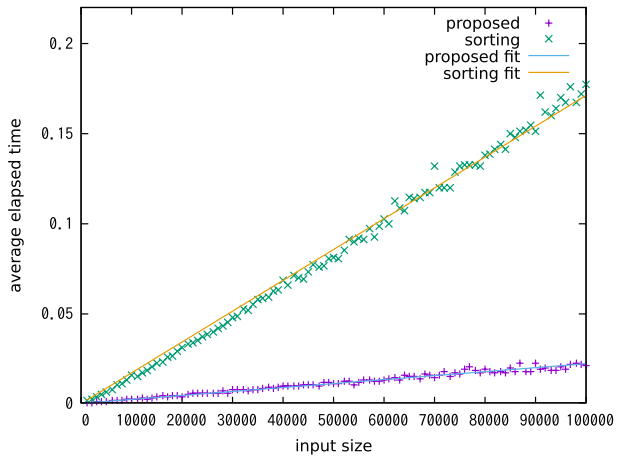

Fig. 9 Performance of the proposed algorithm (plus symbol), horizontal axis for the input data size and vertical axis for averaged elapsed time. Sorting method ("cross" symbol) result is added for comparison, which shows downward convex shape. $n \log(n)$ fits to the sorting method result with residual 7.7e-04, which is significantly better than linear fitting(shown) of 1.6e-03

(published in 1961 as Hoare 1961). Quickselect works on the given rank k, which is not known when h-index is to be calculated.

1. The algorithm takes two arguments, the data array to be processed and additional numeral, which is to keep the temporary value of the index while processing. It is initialized as 0.
2. Create two empty arrays as *upper* and *lower*. Create a numeral *eq* as 0.
3. Pickup a single element from the *data* as *pivot*, compare it with each of the *data*'s elements. If $pivot < element$, push the element into the *upper* array. If $pivot > element$, put it to *lower*. If $pivot = element$, increment *eq* by one and discard the element.
4. If $upper.length + count - pivot$ is equal to 0 or 1, return *pivot* and exit. This is the BINGO situation.
5. Otherwise, if $upper.length + count > pivot$, the h-index should be found in *upper*. Consequently, return to step1 with the argument data as *upper* and the numeral argument *count* is unchanged.
6. Otherwise, the h-index should be found in *lower*. Therefore starts from step 1 with new data as *lower* . As $eq + upper.length$ elements were found above *lower*, the numeral argument must be incremented to $count + eq + upper.length$.

Each time we go back to the beginning of the algorithm, the data to be processed is expected to be half the size[4]. Consequently, comparison and data separation operations will be executed

[4] Regardless of the distribution of the data because the rank is always uniformly distributed.

$$\sum_{i=0}^{\infty} \frac{n}{2^i} = 2n \tag{10}$$

times in total.

Figure 9 shows plot of performance measurement, in which the processing time of the proposed algorithm shows linear response to the input size, as expected from Eq. 10. In comparison, sorting method shows downward convex response, which is again expected from the burden of sort operation. These two methods were tested with identical randomly generated data set of designated size, twenty runs were averaged for each data.

Because the total computational time is given as the sum of computational time necessary to process each segmented data, if segmentation is consistently unbalanced, for example segmented to a single element array and the rest, the total computational time should be $\frac{n(n+1)}{2}$. It means the worst case will take $O(n^2)$ of computational time, which is noted in most of the information science text books (see Press et al. 2007, Chapter 8). $O(n^2)$ of time complexity can not be tolerated if we are to process large scale data.

Fortunately, we virtually have no need to consider this worst case as the computational time converges to $O(n)$ in probability. We will give a brief proof of convergence as follows:

Let a_i be the size of $i-$th segmented data while the algorithm is running. As we select the pivot randomly, the range of a_{i+1} is given as $[0, a_i - 1]$, within this range a_{i+1} is distributed uniformly. Let expected upper bound of each segment size a_i be $R(a_i)$ (their lower bound is always 0, which corresponds to the BINGO situation). Additionally, for any probabilistic variable x, let $E(x)$ be expectation and $Var(x)$ be variance of the variable as usual.

Clearly, $R(a_{i+1}) = E(a_i)$, which is $\frac{R(a_i)}{2}$ as a_i is also distributed within its range. Therefore $Var(a_{i+1}) = \frac{Var(a_i)}{4}$ and the first partition size has variance of $Var(a_1) = \frac{n^2}{4}$. From which the variance of overall computation results as follows: $Var(\sum_i a_i) \leq \sum_i Var(a_i) = \frac{n^2}{3}$.

Let $C(n) = \sum a_i$, the total computational time to process data of size n. From the definition of $O(n)$, $C(n) \neq O(n)$ is rewritten as $\forall \epsilon\ C(n) > \epsilon n$. Because ϵ is arbitrary, we can set it to an unbounded above monotonic function $\epsilon(n)$ and rewrite the condition as $C(n) > \epsilon(n)n$. From the fact that $Var(C(n)) \leq n^2$, probability $P(C(n) > O(n)) \leq \frac{1}{\epsilon(n)}$ concludes directly from Chebyshev's inequality, which implies our claim of convergence in probability of the computational time to $O(n)$.

Above discussion also ensures that we encounter only finite $C(n) > O(n)$ cases even if $n \to \infty$. Actually the probability $P(C(n) = \frac{n(n+1)}{2}))$ is given as $\frac{1}{n!}$, which is less than 1.0e-150 when n is only 1.0e+02; typical data we contemplate has up to 1.0e+08 of size. Actually, as seen from Fig. 9, the proposed algorithm performs more than ten times faster than conventional sorting method even if the data has only 1.0e+05 of size.

Here is a sample implementation of our proposed algorithm in LISP, which is virtually a direct translation from the algorithm description.

```
(defun hirsch (data &optional count)
  (if (not data) count
    (let* ((upper ())
           (lower ())
           (pivot (sample data))
           (eq 0)
           (count (or count 0)))
      (mapcar (lambda (e) (if (< pivot e)
                              (setq upper (cons e upper))
                            (if (> pivot e)
                                (setq lower (cons e lower))
                              (setq eq (+ 1 eq))))) data)
      (if (or (eq pivot (+ 1 count (length upper)))
              (eq pivot (+ count (length upper))))
          pivot
        (if (> (+ count (length upper)) pivot)
            (hirsch upper count)
          (hirsch lower (+ (length upper) eq count)))))))
```

Abbreviations

h-index: A widely recognized and utilized bibliometric index developed by J. E. Hirsch. See (Hirsch 2005); *h5*-index: A derivative of the *h*-index defined on the bibliometric data of five-years of interval; *h*-dimension: A fractal dimension of the *h5*-index or *h*-index, denoted by h_d. The main proposal of this study; EBPM: Evidence-Based Policy Making; CSTI: Council for Science, Technology and Innovation, which is a department of Japan Cabinet Office; e-CSTI: An EBPM system of science policy of Japanese government and related organizations, developed by CSTI.

Acknowledgements

The authors thank Prof. Amane Koizumi of National Institutes of Natural Sciences, Japan for his kind advice, exact and comprehensive explanation of bibliometrics, especially *h*-index. The authors also thank Prof. Takahiro Ueyama, full-time executive member of the CSTI of the Cabinet Office for initiating and promoting of the overall project, invaluable advice and discussion for this study. The authors also thank Mr. Iwao Miyamoto, former director of the Cabinet Office for his direction and management of the overall project, as well as useful discussion and helpful advice for this study. The authors also thank Mr. Toshihiro Shirai, director of the Cabinet Office for his direction and management of the overall project, as well as for the help to publish this study. The authors also thank Professor Naoshiro Shichjo of Graduate Institute for Policy Studies for his effort for the arrangement of additional data acquisition and invaluable advises. The authors also thank Mr. Hiroyuki Sano of mixi, Inc. for preparing additional data and its processing. The authors also thank Professor Miki Haseyama, Mr. Teruo Nishioka, Professor Toshiaki Fujii, Professor Shingo Ebata, Professor Atsushi Kaneda, Dr. Ryoichi Kawada, Mr. Shinsuke Kawachi, Dr. Satoshi Nishiyama, Professor Manami Matsubayashi, Ms. Yoko Inoue, Ms. Minako Ikehata, Mr. Takuya Iwasaki, Mr. Yukio Iwasa, Mr. Yasuhiro Yamaguchi, Mr. Yasuhiro Kubo, Dr. Shinichi Akaike, Mr. Hideyuki Maiwa, Mr. Tomonao Takamatsu, Mr. Hajime Aasano, Mr. Teruki Mita, Dr. Toru Shinohara, Dr. Hideo Arimoto, Mr. Toru Sasabayashi, Mr. Akira Sanagi, Mr. Hiroshi Okumura, Professor Takashi Ishida, Mr. Masato Mizuno for their encouragement and support for this study. Data sourced from Dimensions, an inter-linked research information system provided by Digital Science (https://www.dimensions.ai). Data licensed to Japan Cabinet Office as part of the e-CSTI project.

Authors' contributions

YF is responsible for the design and implementation of the analysis and index calculation algorithm. NU is responsible for the design and implementation of the analysis and data collection. All authors read and approved the final manuscript.

Funding

This study was conducted as a part of regular assignment of the Cabinet Office of the Japanese government. The academic paper data set is acquired by the Japanese government's funding to the Cabinet Office for e-CSTI development.

Availability of data and materials

The data used in this study is an exclusive property of the data service providing company.

Code availability

Reference implementation of the algorithm is available as the "Appendix", as well as in the corresponding author's Github repository: https://github.com/fjt/optimium/blob/master/stat.rb.

Declarations

Competing interests

The authors have no known competing interests with the data source, results and the outcome of this study.

Author details
[1]Japan Cabinet Office, Tokyo 1008914, Japan. [2]Turnstone Research Institute, Inc., Kamakura 2480004, Japan. [3]Nagoya University Graduate School of Engineering, Nagoya 4648603, Japan.

Received: 25 August 2021 Accepted: 14 November 2021
Published online: 06 January 2022

References

Ahlgren P, Sjögårde P (2015) Formal definitions of field normalized citation indicators and their implementation at kth royal institute of technology http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-267026
Alonso S, Cabrerizo FJ, Herrera-Viedma E, Herrera F (2009) h-index: a review focused in its variants, computation and standardization for different scientific fields. J Informetr 4(3):273–289
Barabasi A, Albert R (1999) Emergence of scaling in random networks. Science 286(15):509–512
Bornmann L, Haunschild R (2016) Citation score normalized by cited references (csncr): the introduction of a new citation impact indicator. J Informetr 10(3):875–887. https://doi.org/10.1016/j.joi.2016.07.002
Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C (2013) On the origins of hierarchy in complex networks. Proc Natl Acad Sci 110(33):13316–13321. https://doi.org/10.1073/pnas.1300832110
Data-Science I (2019) The thirty-third public release of grid. https://doi.org/10.6084/m9.figshare.8137970
Dimensions.API.document.team (2020) The official documentation of the dimensions search languages. https://docs.dimensions.ai/dsl/
Elsevier (2019) Research metrics guidebook. https://www.elsevier.com/__data/assets/pdf_file/0020/53327/ELSV-13013-Elsevier-Research-Metrics-Book-r12-WEB.pdf
Hill BM (1975) A simple general approach to the inference about the tail of a distribution. Ann Stat 3:1163–1174. https://doi.org/10.1214/aos/1176343247
Hirsch JE (2005) An index to quantify an individual's scientific research output. PNAS. https://doi.org/10.1073/pnas.0507655102
Hoare CAR (1961) Algorithm 65: find. Commun ACM 4(7):321–322. https://doi.org/10.1145/366622.366647
Koizumi A (2018) Kenkyuuryoku-no-hakarikata (in Japanese). Gakujutsu-no-doukou (in Japanese) 23(12):64–67. https://doi.org/10.5363/tits.23.12_64
Koltun V, Hafner D (2021) The h-index is no longer an effective correlate of scientific reputation. PLoS ONE 16(6):e0253397. https://doi.org/10.1371/journal.pone.0253397
Mandelbrot B (1977) Fractals: form, chance and dimension. W H Freeman and Co
Newman M (2018) Networks Oxford. https://doi.org/10.1093/oso/9780198805090.001.0001
OECD (1996) The knowledge-based economy
Pratelli L, Baccini A, Barabesi L, Marcheselli M (2012) Statistical analysis of the Hirsch index. Scand J Stat 39(4):681–694
Press WH, Teukolski SA, Vetterling WT, Flannery BP (2007) Numerical recipes. Cambridge University Press, Cambridge
Price DDS (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci. https://doi.org/10.1002/asi.4630270505
Qian Y, Rong W, Jiang N, Tang J, Xiong Z (2017) Citation regression analysis of computer science publications in different ranking categories and subfields. Scientometrics. https://doi.org/10.1007/s11192-016-2235-4
Reddy V, Gupta A, White MD, Gupta R, Agarwal P, Prabhu A, Lieber B, Chang YF, Agarwal N (2020) Assessment of the NIH-supported relative citation ratio as a measure of research productivity among 1687 academic neurological surgeons. J Neurosurg 31:1–8. https://doi.org/10.3171/2019.11.JNS192679
Song C, Havlin S, Makse HA (2005) Self-similarity of complex networks. Nature. https://doi.org/10.1038/nature03248
Waltman L (2016) Special section on size-independent indicators in citation analysis. J Informetr 10(2):645. https://doi.org/10.1016/j.joi.2016.04.001
Waltman L, van Eck NJ (2012) The inconsistency of the h-index. J Am Soc Inf Sci Technol. https://doi.org/10.1002/asi.21678
Zhao J, Yu H, Luo JH, Cao ZW, Li YX (2006) Hierarchical modularity of nested bow-ties in metabolic networks. BMC Bioinform 7:1–16

Publisher's Note