Scientometrics (2017) 111:1751–1774
DOI 10.1007/s11192-017-2285-2

The fractal dimension of a citation curve: quantifying an individual's scientific output using the geometry of the entire curve

Antonia Gogoglou[1] · Antonis Sidiropoulos[2] · Dimitrios Katsaros[3] · Yannis Manolopoulos[1]

Received: 26 September 2016 / Published online: 16 February 2017

Abstract Assorted bibliometric indices have been proposed leading to ambiguity in choosing the appropriate metric for evaluation. On the other hand, attempts to fit universal distribution patterns to scientific output have not converged to unified conclusions. To this end, we introduce the concept of *fractal dimension* to further examine the citation curve of an author. The fractal dimension of the citation curve could provide insight in its shape and form, level of skewness and distance from uniformity as well as the existing publishing patterns, without a priori assumptions on the particular citation distribution. It is shown that the notion of fractal dimension is not correlated to other well-known bibliometric indices. Further, a thorough experimentation of the fractal dimension is presented by using a set of 30,000 computer scientists and more than 9 million publications with over 38 million citations. The distinguishing power of the fractal dimension is investigated when comparing the impact of scientists and when trying to identify award winning scientists in their respective fields.

Keywords Fractal dimension · Citation curve · Scientist ranking

✉ Antonia Gogoglou
agogoglou@csd.auth.gr

Antonis Sidiropoulos
asidirop@it.teithe.gr

Dimitrios Katsaros
dkatsar@inf.uth.gr

Yannis Manolopoulos
manolopo@csd.auth.gr

[1] Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

[2] Department of Information Technology, Alexander Technological Educational Institute of Thessaloniki, 57400 Thessaloniki, Greece

[3] Department of Electrical and Computer Engineering, University of Thessaly, 38221 Vólos, Greece

Introduction

During the last decade we have witnessed an explosive growth in the topic of computerized evaluation of the performance of scientists and publications. The result of this growth was the development of a variety of scientometric indicators. From simple measures such as the (average) number of citations (measuring the impact) as captured by the Impact Factor (Garfield 1955), and measures such as the number of papers (measuring the productivity), the community proposed more sophisticated indicators. Examples of these include the indicators which act as a proxy of impact and productivity, as captured by the family of the *h*-index (Hirsch 2005) indicators, centrality indices that measure the strategic position of an element (scientist or journal) within a complex network of citations as captured by the variations of the PageRank (Brin and Page 1998) (e.g., the Scimago[1] and the Eigenfactor[2] indicators) and so on. All these efforts aim towards developing more accurate measures of quality in order to facilitate decisions about faculty promotions, personnel hiring, fund allocation and article submissions, since these tasks have attracted a growing interest from the scientific community (Glänzel et al. 2016). What is more, they have proven to be extremely delicate tasks due to the variety of criteria that need to be considered and also time-consuming due to the large number of candidates (scientists or publications).

At the heart of all these indicators, conventional and novel, we can recognize an effort to summarize the wealth of information carried by a citation curve into a single number. For instance, the total number of citations is the zeroth moment of the citation distribution (i.e. curve), the average citation rate is the first moment of the distribution, the *h*-index is a lower bound of citations, the PageRank-like indicators are the principal eigenvector of a Markovian chain over the (enhanced) citation network, etc. In practice, this effort has been proved challenging, and this is the reason that a number of organizations have decided (Callaway 2016) to move on to publishing the whole citation curve instead of single number summaries.

The present article employs the concept of *fractal dimension* of a point set to allow for a single number indicator that can incorporate geometric information of the citation curve in an accurate and representative manner. This way information about the publishing patterns could be revealed without a priori assumptions on the distribution of citations e.g., being a power law, or a predefined relationship between different parts of the curve. We present the new indicator and results about its performance using citation data of individual scientists; however the methodology developed can be applied to journal citation curves as well.

Motivation and contributions

Consider an author i with p publications and $C_{\max}$ number of citations received by her/his most cited publication. The citation curve is not continuous but rather a set of points; therefore, a curve is fitted to connect them and graphically illustrate the distribution of the number of citations to an author's publications. The area defined by the two axes and the citation curve corresponds to the total number of citations, C_{tot}, acquired by all p publications of author i. Ideally, the maximum *h*-index would occur when the citation

[1] http://www.scimagojr.com/.

[2] http://www.eigenfactor.org/.

curve becomes a straight line parallel to the axes of publications. For the purposes of the present work we refer to this curve as the "maximum citation curve", since it indicates that every publication of a particular author has received citations equal to C_{max}. In practice, this case is rarely met, because there are always publications with fewer citations than the most cited paper. Consequently, as some papers fail to acquire C_{max} citations, the citation curve starts approximating the curves in Fig. 1 indicated as *line 1, line 2*, etc. There comes a point at which the citation curve reduces to *line t* forming a triangle with the two axes. Any citation curve that lies below this line, becomes increasingly skewed indicating the existence of a few highly cited publications, an *h*-core and a long tail with low or zero cited publications. This constitutes the most common case for citation curves.

The more a citation curve differs from the maximum citation curve, the more skewed it becomes. Citation curves significantly different from line *t* and closer to the origin of the axes represent a heavily-tailed and skewed publishing behavior. Therefore, the concept of fractal dimension could be utilized to convey this geometrical information of the citation curve. The fractal dimension provides a statistical index of the complexity of a geometrical object by comparing how much the detail in a pattern changes with the scale at which it is measured. As the citation curve is not in reality a continuous curve but a set of discrete points, the fractal dimension can better represent it than any metric that attempts to quantify parts of the citation curve and the relationship between them. Moreover, it provides an insight on the degree the curve differs

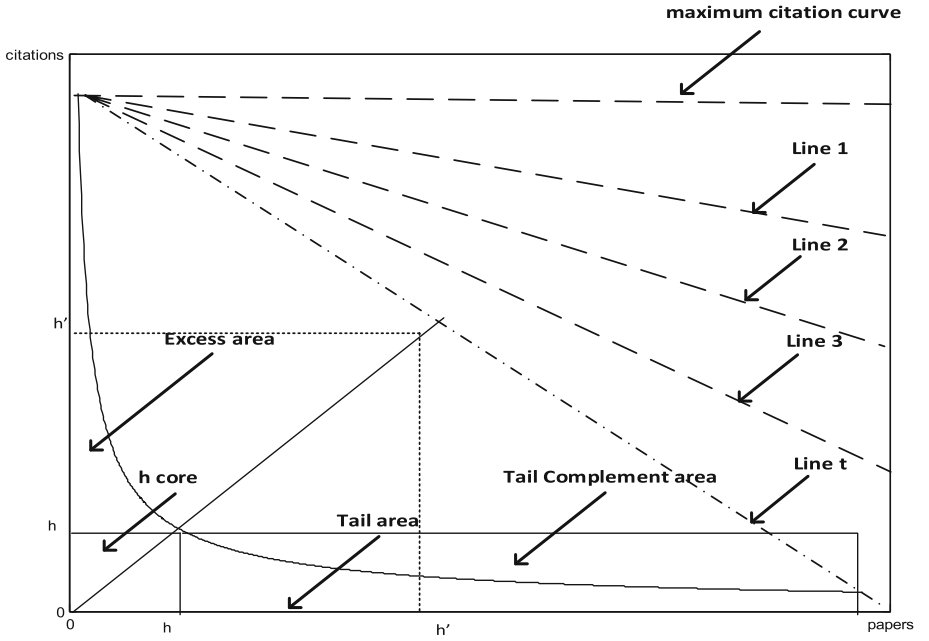

Fig. 1 Citation curves representing different levels of scientific impact with the same number of publications (p) and maximum citations received by individual papers (C_{max})

from a uniform distribution, how self-similar, dynamic and exponential it is through a single number metric. That could prove useful for several data mining tasks performed on bibliometric data (extracting top scientists from a group, ranking, clustering scientists in groups, etc.), because, by utilizing the fractal dimension, information about the geometrical features of the citation curve can be incorporated as numerical input to these tasks. In addition, the fractal dimension could assist in a more reliable distinction amongst authors in very densely populated areas of *h*-index and citation scores (see "Experimentation" section).

Given the above framework, the contribution of this work is twofold. Firstly, given the current state of a scientist (i.e. $p, C_{\max}, C_{\text{tot}}$), the fractal dimension expresses how much this particular state differs from the maximum citation curve. In other words, it indicates with a single number to what extend an author has a productive and highly cited portfolio relative to their career length through measuring the skewness of her/his citation curve. The second contribution of our proposal is revealing the distinguishing power of the fractal dimension especially for common values of p, C_{tot} and *h*-index. To this end, a series of experiments were performed by using a data set of 30,000 authors in a network containing more than 9 million publications and over 38 million citations. More specifically, the fractal dimension calculated by the boxcount method was used to identify the different publishing patterns and scientific impact of each author. Groups of scientists that have received awards in their respective fields were used to explore the distinguishing power of the fractal dimension as a means for identifying elite scientists.

The rest of the paper is organized as follows. "Related work" section summarizes related work; "The fractal dimension and its calculation" section introduces the basic principles of fractal analysis and an overview of the uses of the fractal dimension in the literature. "From power laws to the fractal dimension of the citation curve" section connects the power law nature of the citation curve with fractal dimension, whereas "Experimentation" section presents the experiments conducted to validate the fractal analysis. "Conclusions and future work" section concludes the article.

Related work

A lot of research has focused on studying the process according to which citations are accumulated. There are two different approaches to the analyses of citation dynamics, the *microscopic* and the *macroscopic* one, with the first referring to the study of individual citation curves, whereas the latter refers to the citation distribution across a large set of points (scientists, publications, etc.).

The *microscopic* approaches entail identifying particular characteristics of the citation curve of an author. Either for individual authors (Wildgaard et al. 2014) or for more complex networks of authors (Nykl et al. 2014) the citation curve has been used to quantify scientific productivity and impact. Jorge Hirsh focused on the citation curve and its intersection with a 45° line to define the popular *h*-index (Hirsch 2005). Variations of the *h*-index examined other areas defined by this intersection: the *h*-core area, the *excess* area and the *tail* area. The *h*-core is the square of size *h* defined by the projections of the intersection point on the two axes, whereas the excess area lies between the upper side of the core-square and the citation curve itself. The tail area includes the space on the right of the core-square including the papers with less than *h* citations. The tail-core ratio has been connected to the power law nature of citation distributions (Ye and Rousseau 2010). These

areas have been further used to define indices such as: the *e*-index (Zhang 2009), and the *Perfectionism Index* (*PI* index) (Sidiropoulos et al. 2015) focusing on a new area, the tail complement area. As can be seen in Fig. 1, the tail complement area measures how far are the low cited publications from receiving *h* citations each and contributing to the author's *h*-index.

Among other variations of the *h*-index, we mention the following ones: the *contemporary h*-index (Sidiropoulos et al. 2007), the *individual h*-index (Batista et al. 2006), the *A*- and *R*- indices (Jin et al. 2007), which utilize information carried by the size of the *h*-core. On the other hand, indices like the aforementioned *PI* index exploit the relative size of different areas in the citation curve and try to combine them in one metric. Indices like *s*- or *entropy* index (Silagadze 2010), *w*-index that creates classes for the citation count (Wohlin 2009) and h_m-index (Miller 2006) attempt to incorporate citations from all the areas of the citation curve, thus, tapping into the information the citation curve carries as a whole. However, all these indices do not describe the actual distribution of the citation curve neither do they convey information about the geometry of the entire curve (shape, form, steepness, skewness, etc.). Instead, they break down the curve to individual parts and then attempt to combine them.

From a *macroscopic* point of view citation distributions have been modelled and analyzed multiple times over the years (Stringer et al. 2010; Wallace et al. 2009; Radicchi and Castellano 2012), mainly as a complex network of citations. The dynamics of citation distribution have been analyzed and the possible models that could be fitted to describe them—such as power law, log-normal and shifted power laws—have been introduced in Eom and Fortunato (2011). Analogous work has been conducted in Radicchi et al. (2008) to identify a universal scaling parameter and re-scale citation distributions from different fields on a common universal scale. In Gupta et al. (2005) the authors focused on fitting a Tsallis (power law) model to the distribution of the total citation index over a number of publications, whereas other researchers have focused on adapting similar models to the individual citation curve. A two-phase model has been adapted to describe both the exponential parameter and the power law tail of the citation curve in an attempt to quantify "the rich getting richer" phenomenon in citation distribution (Peterson et al. 2010). However, these attempts are defined at a publication-level and do not provide distinguishing properties at author-level based on an author's entire portfolio. Recently, the power law models for citation distribution have been subjected to scrutiny and their applicability over a range of citation networks has been questioned. An empirical study has been conducted over a very large dataset across different disciplines to compare various long tailed distributions and identify the ones that better describe the citation network, with distributions like Gumbell and Yule law, proving more fitted than power laws (Brzezinski 2015). Moreover, fitting a distribution to the citation data with a satisfying goodness of fit typically requires a large set of data (publications-citations), which is not attainable for individual authors with a limited number of articles published.

The fractal dimension and its calculation

A set of points is considered to be fractal (Gouyet 1996) if it exhibits self-similarity over all scales and deviates from uniformity in a geometrical space. Point sets that exhibit these properties exist often in the real world, such as the curve of a coast-line, the shape of a

cloud, etc. Point sets that display self-similarity present the need for a non-integer dimension value, the fractal dimension. Essentially, it is a ratio providing a statistical index of complexity, comparing how detail in a geometrical pattern changes with the scale at which it is measured. To fully comprehend the concept of the fractal dimension for a real data set, we must first distinguish between the *embedding* and *intrinsic* dimension of a dataset.

Definition 1 The embedding dimension E of a dataset is the dimension of its address space. In other words, it is the number of attributes of the dataset. The dataset can have an embedding dimension lower than the dimension of the space where it's embedded. For instance, a line has an embedding dimension of 1, even if it is represented in a higher dimensional space.

Definition 2 The intrinsic dimension D of a dataset is the dimension of the object represented by the dataset, regardless of the space where it is embedded.

If a dataset actually represents a real Euclidean object, then its intrinsic dimension would be equal to its embedding dimension ($D = E$). As it is often the case, the embedding dimensionality of the dataset hides its actual characteristics and does not provide any real insight into the geometry of the object represented by the dataset, when this object fails to obviously resemble a known Euclidean one. The basic properties of the fractal dimension, which expresses the intrinsic dimensionality of an object, are listed below.

Property 1 *The fractal dimension of a Euclidean object corresponds to its Euclidean dimension and is always an integer.*

A point has fractal dimension of 0, whereas a line has a fractal dimension of 1.

Property 2 *The fractal dimension of a dataset cannot be higher than the embedding dimension.*

The fractal dimension can be calculated both for infinite curves and finite datasets. Various techniques have been contemplated for the calculation of the fractal dimension:

- the boxcount dimension (Feng et al. 1996),
- the correlation dimension (Osborne and Provenzale 1989), and
- the information dimension (Ashkenazy 1999).

The most widely used technique to calculate the fractal dimension of real datasets is the boxcount method, and this is the method employed in the present article.

The *correlation dimension* is calculated for any set x of m points in a D-dimensional space as:

$$\mathrm{DC} = \lim_{m \to \infty} \frac{p}{m^2} \tag{1}$$

where p represents the total number of point pairs which have a distance between them that is less than distance ε. As the number of points tends to infinity, and the distance between them tends to zero, the value of DC approximates the following relationship:

$$N(\varepsilon) \sim \varepsilon^{\mathrm{DC}} \tag{2}$$

The slope of the log–log plot of the differential of DC versus distance ε will yield an estimate of the correlation dimension DC. The values for distance ε are chosen with respect to the density and dimensionality of the set x. However, this approximation works only for

higher dimensional objects and for points that tend to be evenly distributed, thereby constituting the correlation dimension unsuitable for calculating the fractal dimension of the citation curve.

Another way to calculate the fractal dimension is to use the *information dimension* which is defined as:

$$\mathrm{DI} = \lim_{m \to \infty} \frac{S(x_m)}{\log_2 m} \tag{3}$$

where $S(x_m)$ is the entropy of the discrete values of set x in a D-dimensional space. The information dimension is generally lower in value compared to the boxcount dimension and higher in value than the correlation dimension. This method for calculating fractal dimension is more suitable for a higher dimensional space, as it usually assigns the vector a dimension much lower than the space in which it is embedded. Furthermore, its calculation requires a large vector containing enough points, so that the limit can have a finite and unbiased value. For the purposes of calculating the fractal dimension of the citation curve we opt for the boxcount method, which will be presented at length in "The boxcount method" section.

The fractal dimension for real finite datasets has found many applications in several data mining tasks, such as describing complex networks (Zhang et al. 2014), or reducing a dataset's dimensionality by identifying redundant attributes that do not affect the fractal dimension of the dataset (Traina Jr et al. 2010). The property of self-similarity in complex networks was first addressed in Song et al. (2005), where it was proven that they consist of self-repeating patterns on all length scales by dividing the complex system into boxes containing nodes within given sizes, i.e. the boxcount method. Fractal dimension has also been employed to describe sets of points that follow non-uniform distributions and fail to comply with known distributions, such as Gaussian, Zipf, Yule, Tsalis, etc (Faloutsos and Kamel 1997). Next, we will present the boxcount method for calculating the fractal dimension of a real dataset.

The boxcount method

The boxcount method is one of the most commonly used techniques to calculate the fractal dimension of a set x in a Euclidean D-dimensional space R^D. The aforementioned D-dimensional space is divided into cubic grid cells of size r. Let $N(r)$ denote the number of these cells that contain at least one point from the dataset. The boxcount dimension or fractal dimension is defined as:

$$\mathrm{DF} = \lim_{r \to 0} \frac{\log\left(N(r)\right)}{\log\left(1/r\right)}. \tag{4}$$

This definition is applied to fractals of infinite number of points. For real data sets with finite number of points, the slope of the boxcount plot is used to quantify the fractal dimension of the set. In other words, the slope of the log–log plot of $N(r)$ versus r gives the estimate of the fractal dimension of the point set as absolute value. If the point set exhibits self-similarity in the range (r_1, r_2), then the plot is almost a straight line. For a point set with self-similarity the following relationship holds:

$$N(r) = N_o \times r^{-\mathrm{DF}}. \tag{5}$$

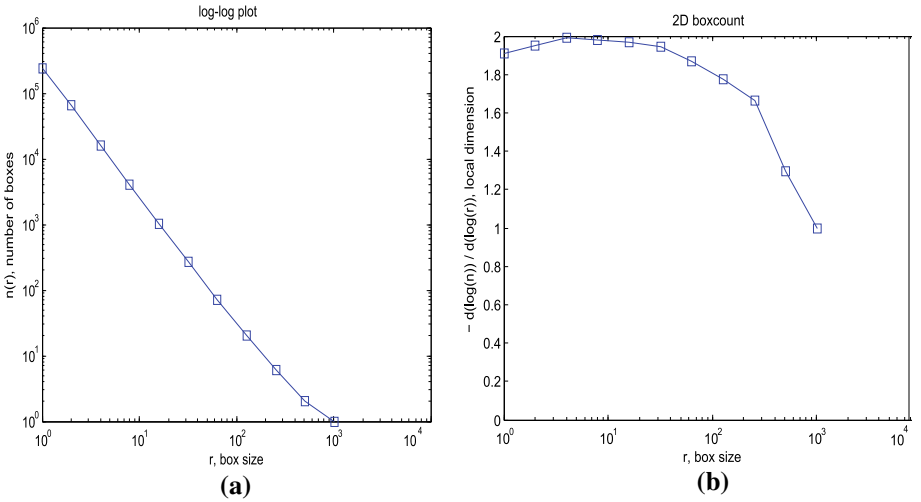

Fig. 2 Log–log and boxcount plots of the Sierpiński triangle. **a** Log–log plot of number of *boxes* versus the size of *each box* for the Sierpiński triangle. **b** Plot of the differential of $\log(N(r))$ to $\log(r)$ versus the size of *boxes* r for the Sierpiński triangle

In Eq. 5 N_o represents the initial box size. In case that the point set significantly differs from a self-similar pattern, then $N(r)$ will represent the highest deviation from the embedding dimension of the point set, i.e. the resulting DF value is the upper limit of fractal dimension for the given set.

Using the Sierpiński triangle, a known object with non-integer fractal dimension that exhibits a self-similar and non-uniform distribution, we plot the number of boxes to the size of the box in doubly logarithmic scales. Figure 2a is the log–log plot for the number of boxes to the size of each box, whereas Fig. 2b depicts the plot of the differential of $\log(N(r))$ to $\log(r)$ versus the size of boxes r, called *the boxcount plot*. The Sierpiński triangle is often used in literature Falconer and Lammering (1998) as a representative example of fractal dimension calculation and as a comparison standard for the expected form of the log–log plot. The slope of its boxcount plot approximates a straight line for the box size in range [1, 100], which comes in accordance to its properties of non-uniformity and self-similarity and based on this slope the Sierpiński triangle's fractal dimension takes the value of 1.75.

The algorithm to calculate the fractal dimension is given below. In particular, the algorithm presented can be used to calculate the fractal dimension of any D-dimensional array C where $D = 1, 2, 3, \ldots d$, for $d \in \Re$. The number N of D-dimensional boxes of size r that are needed to cover the elements of C embedded in a D-dimensional space is calculated for each box size r. The box sizes are powers of 2, $r = 1, 2, 4, 8, \ldots 2^p$, where p is the smallest integer such that $\max(\text{size}(C)) \leq 2^p$ and represents the maximum number of generations of different box sizes. If the size of C over each dimension is smaller that 2^p, then C is padded with zeros to achieve size 2^p over each dimension. Vectors N and r are of size $p + 1$ and N scales as $r^{-\mathrm{DF}}$, based on Eq. 5.

Algorithm 1 Boxcount method for fractal dimension calculation

Require: A D-dimensional array C

Ensure: Fractal dimension DF

Convert C to a bit vector where $C(i) > 0$ equals 1 otherwise 0
width := max(size(C))
p := log(width)/log(2)
Resize vector C so sizes in each of the d dimensions are equal and power of 2
Initialize vector N as $N(1 : p+1) := 0$
$N(p+1) := \sum_{k=1}^{width} C_k$
for j=(p-1) to 0 with step -1 **do**
    $size1 := 2^{p-j}$
    $size2 := round(size1/2)$
    for i=1:size1:(width-size1+1) **do**
        $C(i) := C(i)||C(i+size2)$
    end for
    $N(j+1) := \sum_{k=1,size}^{width-size1+1} C_k$
end for
Reverse vector N
$r := 2^{[0,1,...p]}$
$DFV := -\frac{dlog(N)}{dlog(r)}$
$DF :=$ the average of DFV

Next, we will focus on the relationships of the fractal dimension with the well-known power law and the information it expresses about the citation curve. In the experimental section we will present the boxcount fractal dimension calculated for our dataset and its relationship with well-known metrics, like the *h*-index.

From power laws to the fractal dimension of the citation curve

The calculation of the boxcount dimension assumes that there is a power law relationship between the number of boxes and their respective size. However, this relationship does not require that the dataset itself obeys a pure power law model; it simply implies skewness and deviation from uniformity. It is undeniable though that fractal dimension and power laws are related. In our context we particularly focus on the power law qualities of an individual scientist's citation curve. Recall that a power-law is expressed as follows:

$$p(x) = c \times x^{-\alpha}, \quad (6)$$

where x is the quantity that follows the power-law and α is the *scaling parameter* (or exponent). Constant c is simply a normalization constant.

Even though there are solid studies that show power laws are not a perfect fit for citation distributions (Brzezinski 2015; Garanina and Romanovsky 2016), or that the power law applies after a $x_{\min}$ point of the distribution, the existence of a pure power law is not a prerequisite for the existence and calculation of the fractal dimension of a dataset. Power law-like behaviour is usually detected over large datasets, for instance the citations of all publications in a journal, but the trend towards a power law behavior can be detected even in smaller sets of publications belonging to individual authors (Komulainen 2004).

Using the fractal dimension to describe the shape and form of the citation curve, instead of calculating a power law exponent, is preferable due to the applicability of the fractal dimension over the entire range of points. It can be applied both for scientists displaying heavy tailed distributions and the ones with more uniform patterns. We represent an individual scientist's portfolio as an 1-D vector $C = (c_1, c_2, \ldots, c_m)$, in which $c_i \geq c_j$ if $i < j$ and c_i is equal to the total number of citations accumulated by the i-th publication. Graphically this vector depicts a set of points which, if connected, form a skewed and non-uniform curve, i.e. the citation curve.

As can be deduced, the citation curve lies between a set of individual points and a line, meaning that it constitutes of discrete points which are fitted to a curve to express the distribution of citations. Therefore, it is expected that the fractal dimension of a citation set will lie in the range [0, 1], where values close to 0 mean that the distribution is highly skewed and the points vary from a few high values to a series of zeros. On the other hand, values close to 1 indicate that the citation curve tends to fit a line further away from the origin of the axes and, as it is less skewed, the defined area among the two axes becomes larger resulting in a dense and consistent citation space.

The merit of this approach can be indicated via the following example. Scientist A has 62 publications with a citation vector [22, 20, 19, 15, 15, 14, 14, 13, 11, 11, 10, 10, 8, 7, 6, 6, 6, 6, 6, 5, 5, 5, 5, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] whereas scientist B has a citation vector of [54, 41, 35, 26, 24, 15, 15, 15, 14, 13, 9, 4, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]. It can be easily calculated that both scientists have 271 citations with 62 publications and h-index equal to 10. These conventional bibliometric indices cannot identify differences between the two scientists. It is necessary to examine the whole citation vector and calculate the size of the areas of excess citations or the tail area. If we applied the boxcount method for the two scientists A and B, we could derive their fractal dimension values as 0.890 and 0.671, respectively. Figure 3 shows that the insight provided by the fractal dimension is actually depicted graphically when we plot the citations of a scientist in descending order. Indeed, both scientist A and B start off with a set of well cited papers and they present a set of similarly cited papers in the h-core area. However, scientist B displays a more steep fall and a heavy tailed curve with numerous zero cited papers, whereas scientist A preserves a performance level with a higher number of publications in the h-core area and with several publications receiving close to ten citations each. This observation could indicate that scientist A is closer to raising her/his h-index, thereby displaying higher potential.

The fractal dimension captures the tendency of the citation distribution and provides an estimation of the pattern it would follow if it were to evolve over time. In reality, this is what happens with most scientists' portfolios. They tend to expand dynamically over time

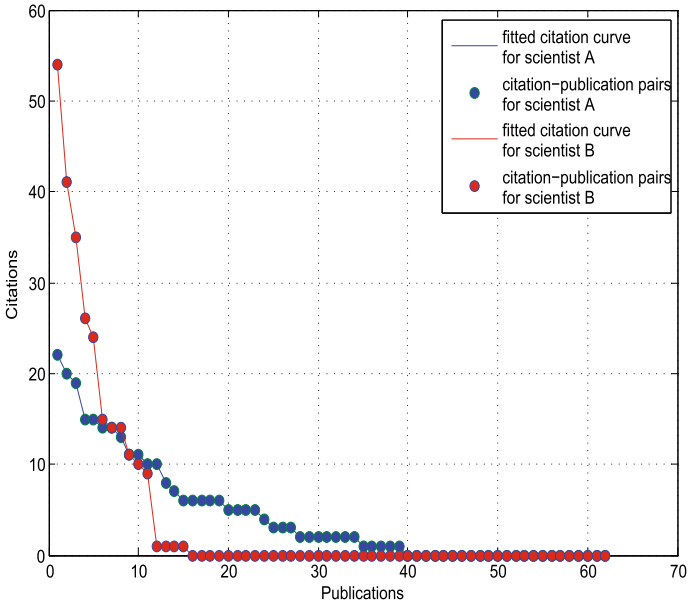

Fig. 3 Citation curves of scientist A displayed with *blue circles* and scientist B with *red ones*. (Color figure online)

and continue acquiring citations following a pattern. As a result, the fractal dimension utilizes the current dataset at the given time to estimate the trend it is going to follow should it continue to expand in a similar pattern. It provides an element of continuity and timelessness in citation distribution and this is why it can assign relatively high values even to scientists with smaller *h*-index values. Since citation acquisition is a dynamic process, it constitutes an oversimplification to focus statically on individual parts of the distribution, as is often the case with most bibliometric studies.

In the next section, more cases of individual authors will be tested to determine the validity of using fractal dimension to describe citation curves and add geometric distinctions amongst a set of authors.

Experimentation

Dataset description

The dataset used in the experiments consists of 30,000 computer scientists according to the categorization of Microsoft Academic Search (MAS) that have an *h*-index higher than 8 as calculated by MAS. The main reason for choosing MAS for data collection is its open access policy and the provision of an API with structured metadata. This allows accessing data for scientists based on the publication field, their *h*-index as calculated by MAS and their number of total publications and citations. Moreover, MAS provides a detailed domain categorization, where each domain is comprised of subdomains that allow for an efficient field specific search. The collected data include information up to the year 2013. The most densely populated time period for the data provided by MAS are the years

1970–2013. The *h*-index threshold of 8 (in the year 2013) was selected to avoid scientists with limited publication count and consequently very small citation curves. The boxcount method can be used for scientists with *h*-index smaller than 8, but the result would introduce bias given the small number of points in the curve resulting in uncertainty in any geometrical estimation about the curve. Another dataset with scientists explicitly acquiring an *h*-index lower than 8 according to MAS was selected for additional experimentation,

In the CS dataset described above we have identified three subsets of award winning scientists in Computer Science in general and the domains of Databases and Networks and Communications in particular:

- the ACM Turing award winners of the years 1980–2015,[3]
- the ACM SIGMOD award winners in the Database domain of the years 1992–2015,[4]
- the ACM SIGCOMM award winners in the Networks and Communications domain of the years 1992–2015.[5]

In addition, we have identified the scientists that have been awarded as ACM Fellows.[6] Out of the 1000 ACM Fellows that are displayed on the ACM website we have extensive publication records for 862 of them in our dataset. It is noted that for a number of the aforementioned award winners not enough data were available in the MAS database, as some of them have had a more industrial profile or made their contributions before the 1970s, a period for which the data in MAS are not as rich. The datasets of the award winning scientists are employed as a comparison set. The values and ranking of the award winning scientists according to the fractal dimension are compared with the ones acquired using other bibliometric indices (such as the *h*-index) to help identify the distinguishing power of the fractal dimension. Table 1 displays the basic statistics of the datasets utilized in the experimental section. The column "CS dataset" contains the full records from MAS for the 30,000 computer scientists, whereas the other four columns refer to the selective datasets of awarded scientists and the last ones contains scientists with *h*-index less than 8.

In Table 1 the average citation rate represents the average number of citations per paper (i.e. citation rate) received by the authors of the dataset and the highest value is achieved by the Turing award winners, who appear to receive on average 158 citations per paper. On the other hand, they also present low publication rate (number of publication per author) compared to the other groups of award winners, which could indicate that Turing award winners produce a small number of seminal and influential publications. Nonetheless, all of the award winning teams display a very high value of fractal dimension (over 0.9) whilst the average *h*-index does not present an analogously high value for all award winning groups. For instance, the SIGCOMM award winners display an average *h*-index value of 36, which is almost half of the respective value for ACM Fellows. The dataset named *LH* contains random scientists from the Computer Science field with an *h*-index lower than 8, as calculated by MAS. This group displays the lowest values in all metrics. These observations and the differences in the spread of fractal dimension compared to traditional metrics will be further investigated in the experimental section below.

[3] http://amturing.acm.org/byyear.cfm.

[4] http://www.sigmod.org/sigmod-awards/.

[5] http://www.sigcomm.org/awards/sigcomm-awards.

[6] http://awards.acm.org/fellow/year.cfm.

Table 1 Statistics of the datasets utilized in our experiments

	CS dataset	SIGMOD winners	SIGCOMM winners	Turing winners	ACM Fellows	LH dataset
Number of authors	30,000	22	33	71	862	350
Number of publications written by authors of the set	2,260,796	3434	5306	9044	277,893	3574
Number of publications including citing papers	9,541,133	120,880	101,022	324,000	11,005,791	10,320
Number of publications per author	75	157	161	127	322	11
Number of citations	38,657,715	304,932	226,937	1,138,584	21,103,115	27,174
Average citation rate of all authors	25	56	50	158	48	8
Average *h*-index	14	52	36	39	60	4
Average fractal dimension	0.75	0.95	0.91	0.92	0.95	0.61

Experimental results

In this section, we present the experiments conducted to validate the use of the fractal dimension. Essentially, the fractal dimension constitutes a normalized metric with values in the range [0, 1]. Table 2 displays the statistics of the fractal dimension for the CS dataset; the values of the fractal dimension are densely centered at an average value of 0.75. It is obvious that the fractal dimension displays a very narrow range of values, as expressed by the small standard deviation and the values of the four quartiles. This comes in accordance to the fact that the citation curve displays a common basic geometry for the majority of scientists. Even though the differences between authors may be considered small in value, the inherent information provides valuable insights. Since bibliometric indices display very different values, any kind of ranking or data mining operation employed on bibliometric data would require normalization to a range of [0, 1] for any index involved in order to facilitate fair comparisons.

Figure 4 displays the empirical cumulative distribution and the fitted distributions to the probability density function of the fractal dimension values of our dataset. As expected, the conclusions drawn from the statistics of Table 2 comply with the distribution plot. Also, it can be seen that the fractal dimension values obey a tailed distribution like most bibliometric indices and the best fitted distribution for it is extreme value or generalized extreme value. This can be explained as fractal dimension values are both upper and lower limited, and concentrated in a range [0.6, 0.9].

We observe that a fractal dimension value equal to 1 (or almost 1) is rather scarce. Such a value reveals that the author has achieved minimum skewness in their citation curve with the majority of their publications receiving high citation count. On the other hand, the latter fact does not necessarily imply that these authors have the highest number of total citations. In other words, high values in the fractal dimension are observed both for moderate citation counts and for highly cited scientists. Consequently, the fractal dimension can prove helpful in distinguishing amongst scientists with similar citation counts or similar *h*-index, particularly in the densely populated groups of scientists with moderate performance. Apparently, values smaller than 0.6 would be common for the fractal dimension

Table 2 Statistics of the fractal dimension for the CS dataset

Mean	SD	Min	Max	1st-quartile	2nd-quartile	3rd-quartile	4th-quartile
0.75	0.108	0.226	1	0.686	0.760	0.828	1

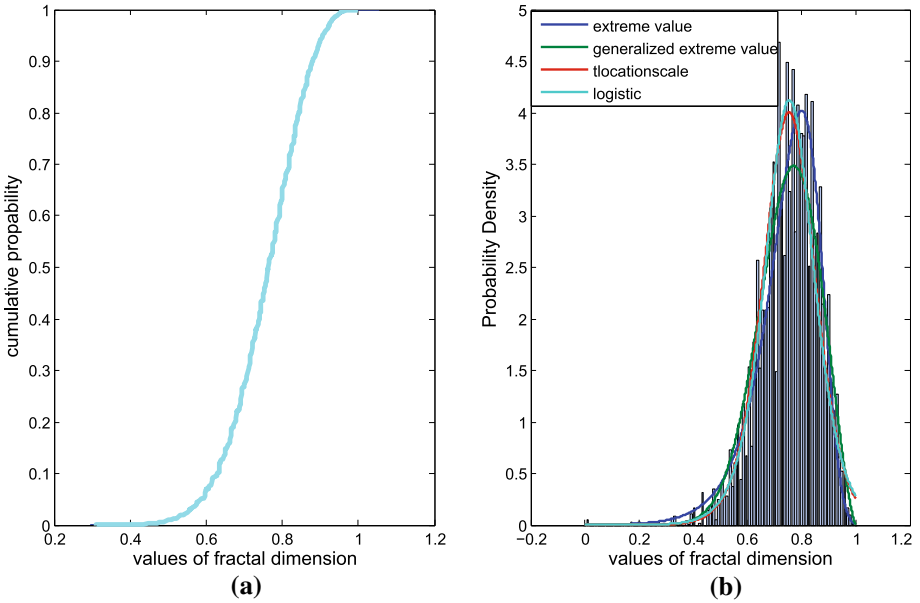

Fig. 4 Distribution of fractal dimension values for the CS dataset. **a** Empirical cumulative distribution function for fractal dimension. **b** Distribution fitted to fractal dimension values

given that many scientists have acquired a low number of citations. But in the CS dataset the authors with very low h-index have been excluded and the behavior of fractal dimension in the lower ranges will be explored further by the end of this section.

Next, we will explore the fractal dimension's relationship with well-known bibliometric indices. Figure 5 depicts the quantile plots of the fractal dimension (y axis) and ten other well-known bibliometric indices (x axis). The indices that were selected for comparison include basic citation metrics, such as the h-index and variations, as well as indices taking into account all the areas around the citation curve. More specifically, the indices contemplated here include: the average number of citations, the total number of citations and the number of publications of an author, the g-index (Egghe 2006), the r_2 index (Gagolewski and Grzegorzewski 2009), the h-index, the h_{nor}-index (Sidiropoulos et al. 2007), the h_w-index (Egghe and Rousseau 2008) and finally the PI and v-index (Riikonen and Vihinen 2008). As can be seen in Fig. 5, the fractal dimension follows a considerably different distribution compared to other bibliometric indices. As a quantile we define the fraction (or percent) of points below the given value. A 45° reference line is also plotted. If the two sets of values follow similar distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the less correlated the values of the two indices.

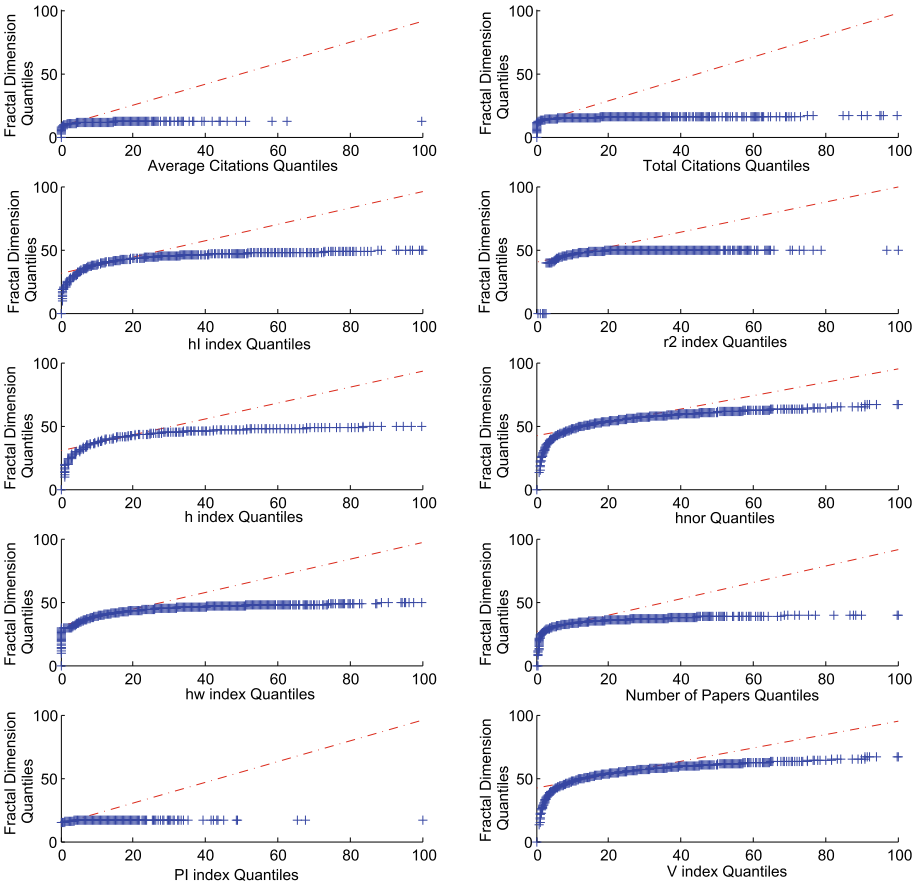

Fig. 5 Collective QQ plots of the fractal dimension with ten other well-known indices

Interestingly, there is a resemblance in the distribution of fractal dimension with the respective one for h_{nor}- and v-index, which is explained because these two indices incorporate information about the entire citation curve. A number of the quantile plots, for instance the ones with g-index, h-index and h_w-index appear to be very similar, as these indices display a high degree of similarity with each other in their distribution. Therefore, compared with fractal dimension, they produce highly similar quantile plots. Other indices that follow a significantly different distribution, such as the PI index, produce a unique quantile relationship with the fractal dimension values. The r_2 index, which has been introduced as a geometrical generalization of various bibliometric indices, appears to follow a more similar distribution to fractal dimension especially in moderate value range; however, for higher values of fractal dimension the quantile distribution deviates significantly from the r_2 values.

Compared with these indices, the fractal dimension follows a rather unique distribution, which we further explore in a temporal context in Figs. 6 and 7. We acquired from MAS the data concerning the same set of authors in earlier years (of course, a number of authors is not included in all the contemplated years depending on academic age) and in particular

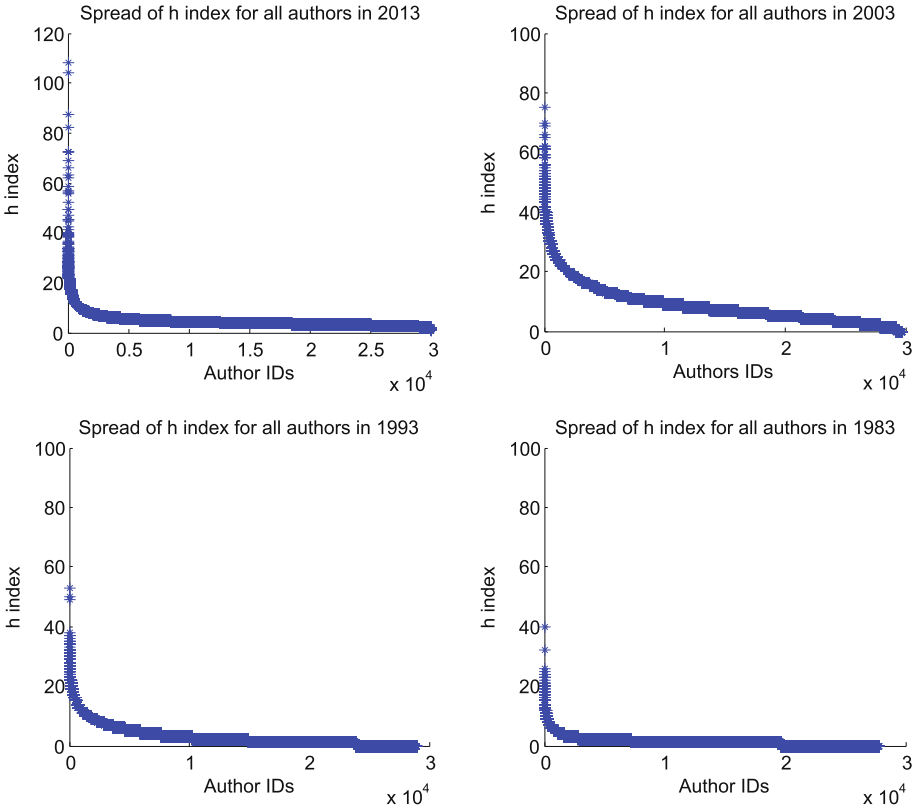

Fig. 6 Spread of *h*-index in years 1983, 1993, 2003 and 2013 over all authors of the Computer Science dataset

the years 2003, 1993 and 1983. In the following, we examine the spread of the fractal dimension and *h*-index over all the authors in our set for these four snapshots in time.

As can be seen in Figs. 6 and 7, both the *h*-index and the fractal dimension get higher values in the more recent years. Although there is an increasing pattern in both metrics, it is observed that the *h*-index has a number of extremely high values, whereas the majority of authors lie in a lower value range [10, 25], which becomes even lower in earlier years. On the other hand, the fractal dimension places a large number of authors in the mediocre value range [0.6, 0.8], which indicates that for scientists of moderate *h*-index values the fractal dimension manages to distinguish a set of scientists that consistently gather citations. In the earlier years examined (i.e. 1983, 1993), since the filtering of *h*-index values higher than 8 was applied only for year 2013, there exist authors with almost zero fractal dimension, meaning there are not sufficient points in their citation curves to apply the boxcount method.

The observations about the spread of the fractal dimension values over time provide an insight in its distinguishing power, especially in identifying high quality scientists for various levels of *h*-index values. In this direction, we have conducted an experiment concerning the distribution of the minimum and maximum value of the fractal dimension for each unique *h*-index value in our dataset. The results are depicted in Fig. 8, where the blue line represents the minimum value of the fractal dimension, whereas the magenta line

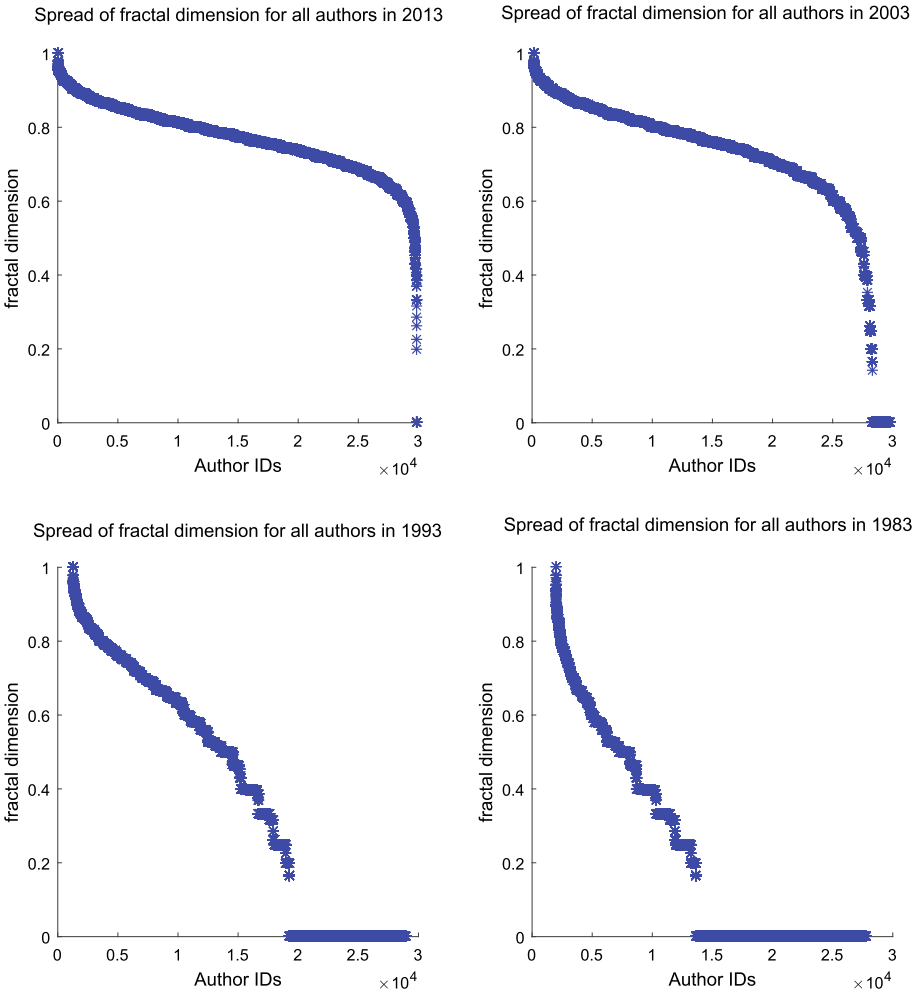

Fig. 7 Spread of the fractal dimension in years 1983, 1993, 2003 and 2013 over all authors of the Computer Science dataset

represents its maximum value. Since these lines are fitted using polynomial regression, they appear to start from the beginning of the axes even though the *h*-index values in our dataset do not include such low values. The circles of different colors represent the award winning scientists of the three mentioned ACM awards (Turing, SIGMOD and SIGCOMM). As can be observed, the award winning scientists have all scored values of fractal dimension close to the maximum, despite the fact they may display various levels of *h*-index values, citation counts, *g*-index values, etc. From our experiments a pattern arises suggesting that distinguished scientists, even if they have not acquired very high values of citation counts in absolute terms, they manage to score high fractal dimension values, meaning they tend to have less skewed citation curves and gather citations consistently following a less heavy-tailed citation distribution.

Another issue arising when exploring how fractal dimension represents the different publishing patterns is the ability to also distinguish moderately performing scientists. In

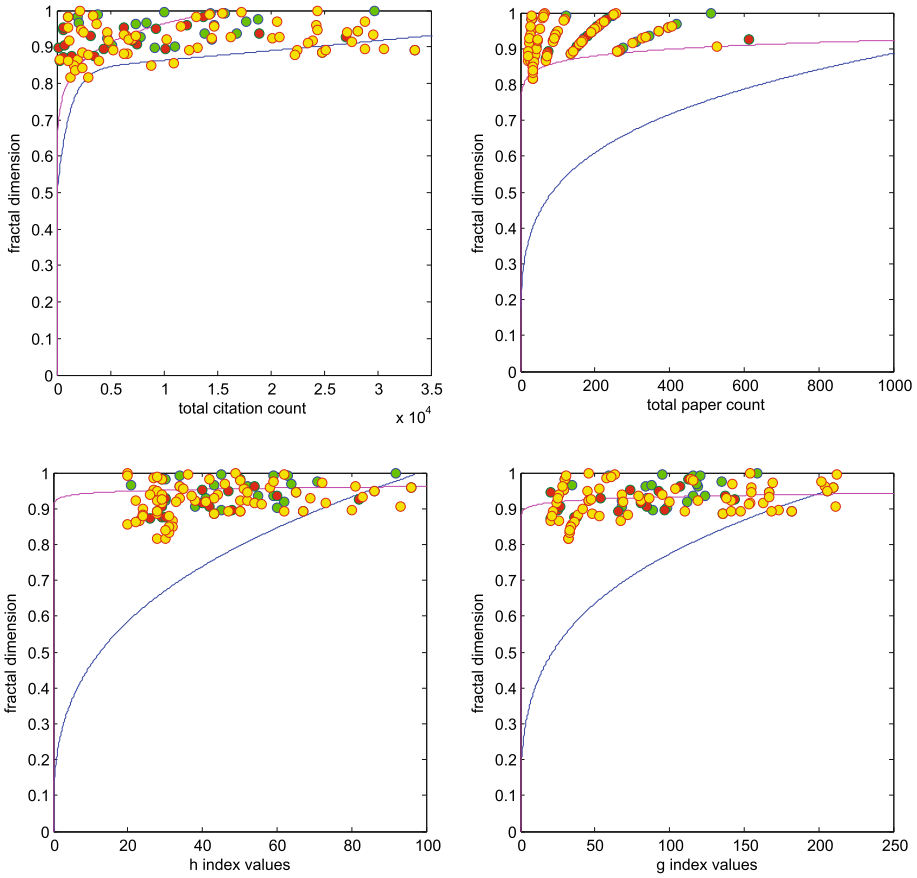

Fig. 8 Curve fitted distribution of the fractal dimension minimum and maximum values for all values of four bibliometric indices: publication, citation count, *h*-index and *g*-index. *Red*, *green* and *yellow circles* represent SIGMOD, SIGCOMM and Turing award winners respectively. (Color figure online)

other words, it is reasonable that top scientists like ACM award winners have a distinguishable value in most bibliometric indices; is it possible though to find similar patterns in well performing scientists who have not been awarded? To answer this question, we have identified the scientists with the highest fractal dimension values in each distinct *h*-index value for the range [26, 50]. The results are displayed in Table 3 where it can be observed that many of the top scientists according to fractal dimension for each *h*-index value are high impact scientists, but have not been awarded with any of the particular prizes. For instance, Victoria Bellotti (CSL/PARC), Roland Chin (Hong Kong University) and André DeHon (University of Pennsylvania) have achieved higher fractal dimension values compared to those of award winning scientists (like David Maier or Donald Knuth) with lower *h*-index values. Analogous examples include Ratul Mahajan (Microsoft Research) and David Dobkin (Princeton University), who have achieved top values in the fractal dimension (>0.99). Surely, award winners of ACM are also included, especially for higher *h*-index values, such as Liskov Barbara and David Maier. From these results, we can deduce that a high *h*-index and high fractal dimension constitutes a pattern for increased

Table 3 Top scientists according to fractal dimension for *h*-index values in the range [26, 50]

Scientist name	*h*-index	Fractal dimension
Rob Glabbeek	26	0.882
Jean-Yves Potvin	27	0.912
Victoria Bellotti	28	0.954
André DeHon	29	0.959
Whang Kyu-Young*	30	0.997
Rudiger Urbanke	31	0.892
Ratul Mahajan	32	0.991
Moshe Tennenholtz*	33	0.971
Jill Mesirov	34	0.979
Tal Rabin	35	0.932
Helmut Boelcskei	37	0.941
Tova Milo*	38	0.963
Jeannette Wing	39	0.936
Margaret Martonosi	40	0.952
David Dobkin	41	0.995
Richard Ladner*	42	0.998
Edward Knightly	43	0.950
Tommi Jaakkola	44	0.973
David Maier*	45	0.927
Gao Lixin*	46	0.996
Donald Knuth*	47	0.943
Saul Greenberg*	48	0.965
Barbara Liskov*	49	0.974
Leslie Valiant*	50	0.960

Scientists with an asterisk have received at least one of the ACM awards

academic impact and complies with the criteria of peer assessment. Further, a high fractal dimension value for moderate citation counts (and *h*-index values) could indicate academic potential and may assist peer decisions in award or grant allocation, tenure committees, etc. It is noted that the most highly populated groups of computer scientists display values of *h*-index between 15 and 35 and it constitutes a real challenge to distinguish a number of high impact scientists in these groups. To this end, fractal dimension may be utilized to distinguish scientists in these densely populated areas based on the geometrical features of their citation curves.

We proceed to investigate the distribution of fractal dimension values in the set of awarded ACM Fellows and discover possible differences in the patterns observed as compared to the general CS dataset. Figure 9 depicts the spread of the fractal dimension and *h*-index values for the 862 scientists that have been awarded as ACM Fellows. As can be seen, compared to Figs. 6 and 7, the spread of fractal dimension values has decreased significantly including only values higher than 0.83, whereas the *h*-index values have not decreased in range. The moderate and high values of *h*-index (>40) are more densely populated but the spread has not changed significantly compared to that of the entire CS dataset. This observation indicates that the fractal dimension is more well adjusted to identify distinguishing scientists compared to traditional metrics.

A more detailed view on the distinguishing ability of the fractal dimension is presented in Table 4, where the top-10 (*group 1*) ACM Fellows that have scored the highest values in

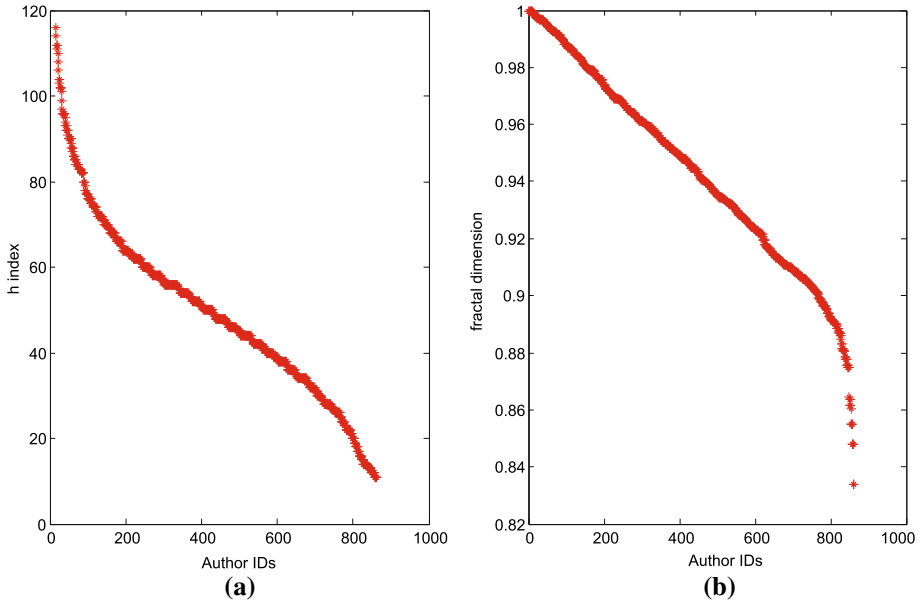

Fig. 9 Spread of *h*-index and fractal dimension for ACM Fellows included in our dataset for year 2013. **a** Spread of *h* index in year 2013 for ACM Fellows. **b** Spread of fractal dimension in year 2013 for ACM Fellows

fractal dimension and the 10 ones (*group 2*) with the lowest fractal dimension value are displayed. Even group 2 displays a fractal dimension higher than the average, but the truly interesting observation is that there exists a wide range of *h*-index values for the ACM Fellows dataset (from 20 to 120), which can be explained based on the different fields of Computer Science each Fellow publishes in and the different time periods during which their work was published (1970–2013). However, for the fractal dimension the values are relatively high for all Fellows, either with high *h*-index values or with lower *h*-index values. Despite the fact that several domains may attract a lower number of citation counts due to their particularity or limited audience, whilst others attract broader interest and a larger number of publications, the fractal dimension can help distinguish high impact publishing behavior across fields. More specifically, in Table 4 we can identify scientists whose seminal work was conducted in earlier decades (1970s) and focuses on fields like compilers, computational algebra and mathematical concepts of computer science, where publications are more scarce but nonetheless seminal. Scientists publishing in these areas, such as Anthony Hearn and Allen Tucker, whose work was mostly mathematical, accumulated a lower *h*-index value compared to other award winning scientists. In these cases, the fractal dimension complies with peer review judgement and distinguishes such scientists from their peers with analogous *h*-index values. In addition, on the top of our list according to fractal dimension are ranked scientists with a long and consistent publishing career. Here, a number of exceptionally high impact scientists can be identified, such as Hector Garcia-Molina, Raghu Ramakrishnan and Paul Dourish. A pattern that arises through the observation of the ACM Fellow dataset is that scientists who continue being active until the end of our chosen time period are more likely to achieve a high fractal dimension. This can be explained, as the addition of new but low cited publications would

Table 4 Top scientists (group 1) and lowest ranking scientists (group 2) according to fractal dimension from the ACM Fellows dataset with their respective *h*-index and fractal dimension values

Author name	*h*-index	Fractal dimension
Group 1		
Hector Garcia-Molina	120	0.999
Raghu Ramakrishnan	75	0.999
Paul Dourish	59	0.999
Barbara Ryder	52	0.998
Richard Ladner	42	0.998
T.V. Lakshman	56	0.998
Lixin Gao	46	0.998
Brad Myers	82	0.997
John Caroll	71	0.997
Whang Kuy-Young	34	0.997
Group 2		
Greg Morrisett	41	0.877
Jack Dennis	30	0.877
Anthony Hearn	24	0.875
Allen Tucker	18	0.875
Harold Stone	26	0.875
Frank Zadeck	22	0.874
Paul Mockapetris	25	0.874
David Wheeler	22	0.874
Kert Akeley	23	0.864
Goyal Ambuij	26	0.863

tend to lower their fractal dimension, but if they manage to continue adding high impact publication their fractal dimension rises, compared to their peers that have stopped publishing. In this sense, the fractal dimension awards productivity but only when it is accompanied by increased impact.

In a different direction, we have also examined the behavior of fractal dimension for lower cited scientists, whose *h*-index is lower than 8 and a portfolio of 11 publications on average. Generally, the smaller the citation curve the less it approximates a line and comes closer to a set of discrete points. As a result, a pattern is difficult to be identified and there may be bias present, since the approximation expressed by Eq. 5 becomes more precise as more points are added (i.e. publications). However, calculating the fractal through the boxcount method allows for more accurate estimations even in less densely populated vectors (smaller portfolios) as compared to the other two methods, correlation and information dimension (see "The fractal dimension and its calculation" section). The quantile distribution of fractal dimension values in this dataset is represented in Table 5. As can be seen, the values differ from the respective ones for the CS dataset (see Table 2) in the sense that smaller values are more common for the low cited group of scientists. However, higher

Table 5 Statistics of the fractal dimension for LH dataset

Mean	SD	Min	Max	1st-quartile	2nd-quartile	3rd-quartile	4th-quartile
0.610	0.227	0.106	1	0.510	0.643	0.846	1

values of the fractal dimension are present in this group as well, which could be a result of the bias introduced due to the small size of the citation vector. It can be deduced that the fractal dimension tends to overestimate the portfolios of low cited scientists, whereas for higher profile scientists it can be more strict comparing to traditional ones, like the *h*-index.

Conclusions and future work

This article considered the issue of using a single numeric indicator to summarize the rich information conveyed by a citation curve. Towards this goal, it proposed the use of the fractal dimension of the curve, and applied the methodology to rank computer scientists.

The study showed that the fractal dimension follows a considerably different distribution compared to other bibliometric indices and shifts the focus towards the geometric properties of the citation curve. In this direction, we investigated the values of the fractal dimension in a large data set of computer scientists and explored its spread and evolution over time as well as its correlation with other well-known bibliometric indices focusing on different properties of a scientist's portfolio. We discovered that the fractal dimension follows a different pattern than widely used bibliometric indices (like the *h*-index) and provides a more detailed segmentation especially in densely populated areas of common values of citation count and/or *h*-index. By comparing the fractal dimension values of scientists awarded by ACM we revealed a pattern indicating that award winning scientists tend to score high values in fractal dimension (higher than 0.9), even when they have not scored distinguishably high *h*-index or total citation count values in absolute terms. Another important finding was that analogous behavior is observed in non-awarded scientists, whose impact though is evident in their fields, distinguishing in this way seminal scientists publishing in less prolific fields.

In its core fractal dimension is a relative metric that expresses how citations are distributed across a scientist's portfolio, with the highest performance being a large portfolio with all highly cited publications. In its calculation, it incorporates (in the size of the boxes) the length of one's career and explores the ratio of zero cited publications to the highly cited ones relative to one's career length and highest citation count. The majority of indexes focus on variations of citation counting; this information is not directly represented by the fractal dimension, which can acquire high values for both high and relatively low citation counts. Should one combine raw citation count and the fractal dimension, a thorough overview of a scientist's portfolio can be provided. As discussed in previous studies (Rubem et al. 2015; Sidiropoulos et al. 2016; Ibez et al. 2016), using more than one bibliometric indexes is required for evaluating scientists. However, care must be taken, so that indexes focusing on different qualities of the scientific impact are chosen, because indexes providing various interpretations of citation counts have been found to be highly correlated (Wildgaard 2015; Bollen et al. 2009), thus leading to biased conclusions when combined to evaluate scientists.

As far as calculation limitations are concerned, the introduced fractal dimension can be calculated for all citation counts or career lengths and no assumption on power law or exponential behavior are required. However, for small portfolios it is challenging to identify clear patterns in the publishing behavior. In that case, fractal dimension will provide an estimation of scientific impact, should it continue to expand in the same way.

We intend to further explore the applicability of the fractal dimension in describing publishing patterns in large publication networks not only at author level, but also for

publishing venues or academic institutions and even for individual publications that have been dynamically acquiring citations over a long period of time. It would be of interest to conduct an analogous analysis for other fields beyond the domain of computer science and compare the results. Furthermore, we are keen to understand the usefulness of the fractal dimension in characterizing citation profiles, which are gaining the interest of scientific community (Chakraborty et al. 2015) and conduct an analysis on how fractal dimension can help model citation curves.

References

Ashkenazy, Y. (1999). The use of generalized information dimension in measuring fractal dimension of time series. *Physica A: Statistical Mechanics and Its Applications*, *271*(3–4), 427–447.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, *68*(1), 179–189.

Bollen, J., van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLOS One*, *4*(6), e6022.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117.

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics*, *103*(1), 213–228.

Callaway, E. (2016). Beat it, impact factor! Publishing elite turns against controversial metric. *Nature*, *535*(7611), 210–211.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2015). On the categorization of scientific citation profiles in computer science. *Communications of the ACM*, *58*(9), 82–90.

Egghe, L. (2006). Theory and practice of the *g*-index. *Scientometrics*, *69*(1), 131–152.

Egghe, L., & Rousseau, R. (2008). An *h*-index weighted by citation impact. *Information Processing & Management*, *44*(2), 770–780.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS One*, *6*(9), e24,926.

Falconer, K. J., & Lammering, B. (1998). Fractal properties of generalized Sierpiński triangles. *Fractals*, *6*(1), 31–41.

Faloutsos, C., & Kamel, I. (1997). Relaxing the uniformity and independence assumptions using the concept of fractal dimension. *Journal of Computer and System Sciences*, *55*(2), 229–240.

Feng, J., Lin, W. C., & Chen, C. T. (1996). Fractional box-counting approach to fractal dimension estimation. In *Proceedings 13th International Conference on Pattern Recognition (ICPR)* (vol. 2, pp. 854–858).

Gagolewski, M., & Grzegorzewski, P. (2009). A geometric approach to the construction of scientific impact indices. *Scientometrics*, *81*(3), 617.

Garanina, O. S., & Romanovsky, M. Y. (2016). Citation distribution of individual scientist: approximations of stretch exponential distribution with power law tails. arxiv:1605.03741

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*, 108–111.

Glänzel, W., Beck, R., Milzow, K., Slipersæter, S., Tóth, G., Kołodziejski, M., et al. (2016). Data collection and use in research funding and performing organisations. General outlines and first results of a project launched by Science Europe. *Scientometrics*, *106*(2), 825–835.

Gouyet, J. F. (1996). *Physics and fractal structures*. Berlin: Springer.

Gupta, H. M., Campanha, J. R., & Pesce, R. A. G. (2005). Power-law distributions for the citation index of scientific publications and scientists. *Brazilian Journal of Physics*, *35*, 981–986.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16,569–16,572.

Ibez, A., Armaanzas, R., Bielza, C., & Larraaga, P. (2016). Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indices. *Journal of the Association for Information Science and Technology*, *67*(7), 1703–1721.

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The *R*- and *AR*-indices: Complementing the *h*-index. *Chinese Science Bulletin*, *52*(6), 855–863.

Komulainen, T. (2004). Self-similarity and power laws. In H. Hyötyniemi (Ed.), Complex systems-science on the Edge of Chaos (vol. 145, pp. 109–122). http://neocybernetics.com/report145/chapter10.pdf

Miller, C. W. (2006). Superiority of the *h*-index over the Impact Factor for physics. http://arxiv.org/pdf/physics/0608183.pdf

Nykl, M., Jezek, K., Fiala, D., & Dostál, M. (2014). Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, *8*(3), 683–692.

Osborne, A. R., & Provenzale, A. (1989). Finite correlation dimension for stochastic systems with power-law spectra. *Physica D: Nonlinear Phenomena*, *35*(3), 357–381.

Peterson, G. J., Pressé, S., & Dill, K. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences*, *107*(37), 16,023–16,027.

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS One*, *7*(3), 1–9.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17,268–17,272.

Riikonen, P., & Vihinen, M. (2008). National research contributions: A case study on Finnish biomedical research. *Scientometrics*, *77*(2), 207–222.

Rubem, A. P. S., de Moura, A. L., & Soares de Mello, J. B. (2015). Comparative analysis of some individual bibliometric indices when applied to groups of researchers. *Scientometrics*, *102*(1), 1019–1035.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch *h*-index for disclosing latent facts in citation networks. *Scientometrics*, *72*(2), 253–280.

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2015). Identification of influential scientists vs. mass producers by the Perfectionism index. *Scientometrics*, *103*(1), 1–31.

Sidiropoulos, A., Gogoglou, A., Katsaros, D., & Manolopoulos, Y. (2016). Gazing at the skyline for star scientists. *Journal of Informetrics*, *10*(3), 789–813.

Silagadze, Z. K. (2010). Citation entropy and research impact estimation. *Acta Physica Polonica B*, *41*(11), 2325–2333.

Song, C., Havlin, S., & Makse, H. (2005). Self-similarity of complex networks. *Nature*, *433*(7024), 392–395.

Stringer, M. J., Sales-Pardo, M., & Nunes, L. A. (2010). Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. *Journal of the American Society for Information Science and Technology*, *61*(7), 1377–1385.

Traina, C, Jr., Traina, A. J. M., Wu, L., & Faloutsos, C. (2010). Fast feature selection using fractal dimension. *Journal on Information and Database Management*, *1*(1), 3–16.

Wallace, M. L., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303.

Wildgaard, L. (2015). A comparison of 17 author-level bibliometric indicators for researchers in astronomy, environmental science, philosophy and public health in web of science and google scholar. *Scientometrics*, *104*(3), 873–906.

Wildgaard, L., Schneider, J. W., & Larsen, B. (2014). A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, *101*(1), 125–158.

Wohlin, C. (2009). A new index for the citation curve of researchers. *Scientometrics*, *81*(2), 521–533.

Ye, F., & Rousseau, R. (2010). Probing the *h*-core: An investigation of the tail-core ratio for rank distributions. *Scientometrics*, *84*(2), 431–439.

Zhang, C. T. (2009). The *e*-index, complementing the *h*-index for excess citations. *PLoS One*, *4*(5), e5429.

Zhang, H., Hu, Y., Lan, X., Mahadevan, S., & Deng, Y. (2014). Fuzzy fractal dimension of complex networks. *Applied Soft Computing*, *25*, 514–518.