Quantifying an individual's scientific output using the fractal dimension of the whole citation curve

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

[1]Department of Informatics, Aristotle University of Thessaloniki, Greece
[2]Department of Electrical & Computer Engineering, University of Thessaly, Greece

Abstract

We propose the use of the fractal dimension of a scientist's citation curve as a performance indicator that can capture its geometric shape, and lead to a measure representative of the whole curve.

I. Introduction

The recent surge of efforts for developing indicators that can quantify the performance of scientists which are subsequently used for allocating funds, deciding promotions and so on, has resulted in several hundreds of such indicators that usually estimate a gross statistic over the citation curve. For instance, the h-index [1] is a lower bound of the area under the citation curve. Even the indicators for journal evaluation such as the Impact factor are based on a plain statistic, i.e., the average citation rate. The inadequacy of this approach has been recently highlighted and proposals for indicators that are representative of the whole citation curve, such as the median of citation counts [2], have been proposed.

Motivated by this shortcoming of current indicators, we propose here the use of the *fractal dimension* of the citation curve as a performance indicator which manages to encapsulate more information about the shape and properties of the citation curve as compared to other approaches.

II. Definition of the fractal dimension of a point set

A set of points is considered to be fractal 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 cannot be fitted to a Euclidean object but tend to follow a dynamic pattern, that given enough points displays self-similarity, present the need for another form of non-integer dimension, the fractal dimension, which constitutes a ratio, providing a statistical index of complexity comparing how detail in a geometrical pattern changes with the scale at which it is measured. A fractal dimension does not have to be an integer. To comprehend the concept of the fractal dimension for a real data set, we must first identify the differences 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.

The basic properties of the fractal dimension are listed below.

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

For instance 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 estimation of the fractal dimension:

- the *boxcount dimension* [3],
- the *correlation dimension* [4],
- the *information dimension*[5].

The most widely used technique to calculate the fractal dimension of real datasets is the boxcount method, and it is the one we have opted for in this article. For the sake of brevity we will not delve into the details of how to calculate the fractal dimension of a point set, but we will provide an overview of a small subset of the results we have obtained from our analysis of this new performance indicator.

III. Applying the fractal dimension to a citation curve

The dataset used in the experiments consists of $30,000$ computer scientists based on the categorization of Microsoft Academic Search (MAS) that have an h-index higher than 8 as calculated by MAS. 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.

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

- the ACM Turing award winners of the years 1980-2015.
- the ACM SIGMOD award winners in the Database domain of the years 1992-2015.
- the ACM SIGCOMM award winners in the Networks & Communications domain of the years 1992-2015.

In addition, we have identified the scientists that have been awarded as ACM Fellows. 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, meaning that 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.

In this extended abstract, we will defer from presenting the details concerning the statistical properties of the fractal dimension in our datasets, and the analysis of its correlation with other indicators; instead, we will focus on displaying the distinguishing power of the fractal dimension for a set of high impact scientists and its the ability to also distinguish moderately performing scientists with academic potential.

Towards this goal, 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 I 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 aforementioned 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 academic impact and complies with the criteria of peer assessment. Moreover, 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

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
Liskov Barbara*	49	0.974
Leslie Valiant*	50	0.960

TABLE I

TOP SCIENTISTS ACCORDING TO FRACTAL DIMENSION FOR h-INDEX VALUES IN THE RANGE $[26, 50]$. SCIENTISTS WITH AN ASTERISK HAVE RECEIVED AT LEAST ONE OF THE ACM AWARDS.

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.

A more detailed view on the distinguishing ability of the fractal dimension is presented in Table II, where the top-10 (*group 1*) ACM Fellows that have scored the highest values in fractal dimension and the 10 ones (*group2*) with the lowest fractal dimension value are displayed. Even the scientists in group 2 display 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 II we are able to 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.

Author Name	h-index	fractal dimension
group 1		
Garcia-Molina Hector	120	0.999
Ramakrishnan Raghu	75	0.999
Dourish Paul	59	0.999
Ryder Barbara	52	0.998
Ladner Richard	42	0.998
Lakshman T.V.	56	0.998
Gao Lixin	46	0.998
Myers Brad	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
Zadeck Frank	22	0.874
Mockapetris Paul	25	0.874
Wheeler David	22	0.874
Akeley Kert	23	0.864
Goyal Ambuij	26	0.863

TABLE II
h-INDEX AND FRACTAL DIMENSION VALUES FOR ACM FELLOWS WITH THE HIGHEST FRACTAL DIMENSION VALUES (GROUP 1) AND LOWEST VALUES (GROUP 2).

IV. CONCLUSIONS

We proposed the use of the fractal dimension of the citation curve as a scientometric indicator to quantify the performance of a scientist, and contemplated its ability to distinguish highly performing individuals in consistency with peer review judgement.

REFERENCES

[1] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[2] E. Callaway, "Beat it, impact factor! publishing elite turns against controversial metric," *Nature*, vol. 535, no. 7611, pp. 210–211, 2016.

[3] J. Feng, W.-C. Lin, and C.-T. Chen, "Fractional box-counting approach to fractal dimension estimation," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*, vol. 2, 1996, pp. 854–858.

[4] A. Osborne and A. Provenzale, "Finite correlation dimension for stochastic systems with power-law spectra," *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 357–381, 1989.

[5] Y. Ashkenazy, "The use of generalized information dimension in measuring fractal dimension of time series," *Physica A: Statistical Mechanics & its Applications*, vol. 271, no. 3–4, pp. 427–447, 1999.