

Mathematical theories of citation

Peer-reviewed author version

EGGHE, Leo (1998) Mathematical theories of citation. In: Scientometrics, 43(1). p. 57-62.

DOI: 10.1007/BF02458394

Handle: <http://hdl.handle.net/1942/796>

Mathematical theories of citation

by

L.Egghe

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium^(*)
e-mail : legghe@luc.ac.be

and

UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium.

ABSTRACT

The paper focusses on possible mathematical theories of citation and on the intrinsic problems related to it. It sheds light on aspects of mathematical complexity as e.g. encountered in fractal theory and Mandelbrot's law. There is also a discussion on dynamical aspects of citation theory as reflected in evolutions of journal rankings, centres of gravity or of the set of source journals. Some comments are given in this connection on growth and obsolescence.

(*) Permanent address.

1. INTRODUCTION.

Citation analysis, as part of scientometrics and informetrics, shares virtually all properties and problems related to these sciences. Hence it is certainly not the purpose of this comment paper to discuss these aspects in detail. The author indeed presupposes that the reader is familiar with informetric and scientometric studies and theories and hence that he/she is able to produce some examples of these common interests.

Therefore, I will - apart from a discussion on general standards (below) - restrict my attention to the study of the typicalities of a possible theory of citation, i.e. on those aspects that distinguish the theory of citation from other scientometric theories. The reader, knowing my background, will not be surprised that I will concentrate my study on possible mathematical theories of citation and on the problems related to this.

Let me, before doing this, stress one important problem that has also been encountered in the paper of Leydesdorff, which we discuss here, Leydesdorff (1998): **standards**. This is not typically a problem in citation analysis, as has been enlightened by Glänzel and Schoepflin (1994), but is so important that it needs to be mentioned here. In the Leydesdorff paper one finds terms as: complexity, dynamics, dimension, network, reflexive theory, eigenstructure, eigenvalue, eigenfrequencies, ... and these terms have been used several times. Each of these key words have a specific meaning in mathematics and it is the author's hope that their use in Leydesdorff's article is within the same scope of definition.

The first four of these terms will be discussed in detail in the sequel. As to the "eigen-" key words mentioned above, I am not sure to which matrix one is referring. I suppose Leydesdorff thinks here of citation matrices of all types. In any case the terms "eigenstructure" and "eigenfrequency" should be explained and put in connection with better known terms such as "eigenvalue" and "eigenvector".

I will now discuss the following topics : complexity and dynamics of theories of citation.

2. COMPLEXITY.

If one wants to discuss complexity one has to indicate the exact object (the machine) on which one wants to work. It seems logical to work in the framework of sources and items : sources are the objects that produce and items are the objects that are produced. In citation analysis this means : sources are articles and items are references or citations. In the first case one speaks of synchronous studies of citation and in the second case of diachronous studies (for the many "well-known" topics from informetrics and scientometrics that are used here without giving the definition, we refer the reader to Egghe and Rousseau (1990)). Of course this machinery is not typical for citation studies : such processes - in Egghe and Rousseau (1990) they are called Information Production Processes, abbreviated

IPPs - also occur in other areas: texts in linguistics (words are sources - also called types - and their use in texts are items - also called tokens), bibliographies (journals or authors are sources and articles are items), in libraries (books are sources and their borrowings are items), in econometrics (employees are sources and their work or salaries are items), and so on. But the framework of IPPs, as applied to citation analysis, will reveal typicalities of a citation theory. Let us go into this in more detail.

First of all, it is well-known that IPPs are a good medium to study complexity in the terminology of fractal theory. We refer to Mandelbrot (1977) or to the more readable interpretation of it in Egghe and Rousseau (1990) for the argument that $1/\beta$, where β is the exponent in Mandelbrot's law is the fractal dimension of the IPP. It is furthermore well-known that the fractal dimension of a fractal is a measure of its complexity. It is clear that a citation IPP will show (much) more complexity in terms of fractal dimensions than "ordinary" IPPs such as bibliographies. Indeed in the latter, the sources are the journals and the items are the articles; in the former one starts with the numerous articles as sources and one studies references or citations as items. If one studies concepts as co-citations or bibliographic coupling the complexity is even more increased (in an intuitive setting: in a quadratic way), which will again be reflected in a substantial increase of the fractal dimension. Hence the reader should be convinced that a fractal theory of citation is important. This idea is also supported by Van Raan (1991).

Citation IPPs can also be distinguished from bibliographies in the sense that the items can have more than one source. Indeed, an article can be used several times as a reference (the citing articles (sources) are hence bibliographically coupled) and the same with the (dual) co-citation case. This is not the case when one considers the journal - article relationship: here an article is published in only one journal. Of course, the citation IPP is in this case comparable with the IPP where one considers authors as sources and publications as items: also here an item can have several sources. In Egghe (1994) these more complex IPPs in which items can have several sources are studied by using the "simple" IPPs (of single source items) and applying convolutions operations on the underlying distributions in order to reach the underlying distributions of the complex IPP. For an inventory on applications of convolution theory in scientometrics and informetrics the reader is referred to Rousseau (1998). Note also the application of convolutions in the study of the influence of publication delays on the obsolescence of a discipline, as given in Egghe and Rousseau (1998). Hence convolutions constitute an important tool in the elaboration of a theory of citation.

This section on complexity is closed by making some remarks on the possible degrees of complexity in a theory of citation. Such degrees are also mentioned in the Leydesdorff paper where he is dealing with the different units of analysis that are possible. It is known to me that it is very important to determine the used units of analysis in a very clear way and at the beginning of each study since it is of influence on the calculated measures (such as the average impact factor IF, Price Index, obsolescence and so on, to give just a few examples in citation analysis). Let us concentrate e.g. on the impact factor. The average impact factor of a field (as composed of the articles in this field) is different from the same but where the field is now considered as composed of journals. One talks in this connection about the

“Global Impact Factor” (GIF) in the first case and about the “Average Impact Factor” (AIF) in the second case. See for this the articles Egghe and Rousseau (1996 a,b).

3. DYNAMICAL ASPECTS.

Together with the concept of complexity, the topic of dynamics is mentioned many times in the Leydesdorff article. It is right to do so: dynamical properties of systems in informetrics and scientometrics are amongst the most important features that exist. Indeed, time evolutionary aspects are adding an extra dimension to any “static” study. Because of the above mentioned high degree of complexity of citation IPPs it is clear that the study of the dynamical aspects is far from trivial. In this section it is the purpose to show the reader some aspects of it, thereby also mentioning the intricate problems that are still left in this domain.

Dynamics of general IPPs have been studied e.g. in Egghe and Rousseau (1996c) in the framework of stochastic processes (over time). The processes that are encountered there are the so-called martingales, submartingales and supermartingales (well-known from the theory of gambling). In this sense the found models are also applicable to the theory of citation.

However, recently, I have studied some “typical” citation problems in the area of the evolution of the set of source journals. Source journals are journals that are selected by the Institute for Scientific Information (ISI) to form the basis of their products such as the citation indexes and the Journal Citation Reports (JCR). One can wonder how such a set evolves in time: indeed the performance of journals, as measured by their citation impact in the world, is variable from year to year and hence so is the set of source journals. One can even generalise this problem by posing the following question (as was done in Rousseau and Spinak (1996)). Suppose one starts with an arbitrary set of “initial source journals” at a certain time (e.g. composed in a country different from the USA - say a developing country where possibly also locally important journals are included) and one checks this set every year hereby using performance criteria similar (or the same) as the ones used by ISI. Will we, eventually, end up with the same set of source journals as is used nowadays by ISI ? As found out in the theoretical study Egghe (1998a) this problem can only be dealt with if we are able to study general stochastic processes in infinite dimensional Banach spaces: we encountered quasi-martingales with values in the Hilbert space L^2 . A similar paper is Egghe (1998b) where martingales-in-the-limit (mils) are encountered. It is the first time that these processes are applied outside mathematics ! This shows the very intricate nature of dynamics of citation processes. For those interested in the mathematical theory of such general stochastic processes, we refer the reader to Egghe (1984) and Edgar and Sucheston (1992). I predict many more applications of this mathematical theory in scientometrics and informetrics.

Another aspect related to dynamical problems of citation analysis is the evolution of the rankings of journals, e.g. according to their impact factors (probably restricted to a subject - cf. the Subject Category Listings in the JCR of ISI). In fact, this problem is somewhat related to the above one on the evolution of a set of source journals,

certainly if one considers all journals, then rank them according to a certain impact measure and then select source journals according to their ranking in this list. A theory of these rankings is far from reality; a first attempt can be found in Nieuwenhuysen and Rousseau (1988).

A very intricate problem is the study of the evolution of networks. The networks we are talking about here (and also in Leydesdorff's article) are all directed graphs (also called digraphs) in the sense that they consist of edges and vertices and where the edges have one direction - typical in all kinds of citation analyses. I cannot predict a possible mathematical solution to this problem. A simplification of this problem lies in the reduction of this problem to the study of the centres of gravity of these networks. Aspects of this can be studied in Egghe and Rousseau (1990) and Rousseau (1989a,b).

Finally, I want to mention the important dynamical topics of growth and obsolescence and their interrelations. Obsolescence (synchronous or diachronous) is certainly a subject that is part of citation analysis. I mention growth here since its study is very similar: the same definitions can be given for the growth rate as well as the ageing rate and similar distributions apply though, of course, with other parameters in order to be able to deal with both phenomena - the simplest example being offered by the exponential distribution $f(t) = c \cdot a^t$, where $a > 1$ in the case of growth and where $0 < a < 1$ in the case of obsolescence. Furthermore there is a rather intricate influence of growth on obsolescence in the following sense. If literature grows, there are more articles that can cite the past, yet, as time passes, there are also more articles available as candidates for a citation. The real effect of growth on obsolescence is not very clear. In Egghe (1993) and Egghe, Rao and Rousseau (1995) this influence has been studied again using the technique of convolutions (cf. *supra*), but different results are obtained for synchronous and diachronous obsolescence.

Bibliography

G.A. Edgar and L. Sucheston (1992), Stopping times and directed processes. Cambridge University Press, Cambridge, UK.

L. Egghe (1984), Stopping time techniques for analysts and probabilists. London Mathematical Society Lecture Notes Series 100, Cambridge University Press, Cambridge, UK.

L. Egghe (1993), On the influence of growth on obsolescence. *Scientometrics*, 27 (2), 195 - 214.

L. Egghe (1994), Special features of the author-publication relationship and a new explanation of Lotka's law based on convolution theory. *Journal of the American Society for Information Science*, 45 (6), 422 - 427.

L. Egghe (1998a), The evolution of core collections can be described via Banach space valued stochastic processes. Preprint.

L. Egghe (1998b), Dynamics of a field list of internationally visible journals: a stochastic model. Preprint.

L. Egghe, I.K. Ravichandra Rao and R. Rousseau (1995), On the influence of production on utilization functions: obsolescence or increased use ? *Scientometrics* 34 (2), 285 - 315.

L. Egghe and R. Rousseau (1990), Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam.

L. Egghe and R. Rousseau (1996a), Average and global impact of a set of journals. *Scientometrics*, 36 (1), 97 - 107.

L. Egghe and R. Rousseau (1996b), Averaging and globalising quotients of informetric and scientometrics data. *Journal of Information Science*, 22 (3), 165 - 170.

L. Egghe and R. Rousseau (1996c), Stochastic processes determined by a general success-breeds-success principle. *Mathematical and Computer Modelling*, 23(4), 93 - 104.

L. Egghe and R. Rousseau (1998), The influence of publication delays on the observed aging distribution of scientific literature. Preprint.

W. Glänzel and U. Schoepflin (1994), Little scientometrics, big scientometrics ... and beyond. *Scientometrics*, 30 (2 - 3), 375 - 384.

L. Leydesdorff (1998), Theories of citation ? *Scientometrics*, this volume.

B. Mandelbrot (1977), *The fractal Geometry of Nature*. Freeman, New York.

P. Nieuwenhuysen and R. Rousseau (1988), A quick and easy method to estimate the random effect on citation measures. *Scientometrics*, 13, 45 - 52.

R. Rousseau (1989a), Kinematical statistics of scientific output. Part I: geographical approach. *Revue Française de Bibliométrie*, 4, 50 - 64.

R. Rousseau (1989b), Kinematical statistics of scientific output. Part II: standardized polygonal approach. *Revue Française de Bibliométrie*, 4, 65 - 77.

R. Rousseau (1998), *Convolutions and their applications in information science*. Preprint.

R. Rousseau and E. Spinak (1996), Do a field list of internationally visible journals and their journal impact factors depend on the initial set of journals ? A research proposal. *Journal of Documentation*, 52 (4), 449 - 456.

A.F.J. Van Raan (1991), Fractal geometry of information space as represented by co-citation clustering. *Scientometrics*, 20 (3), 439 - 449.