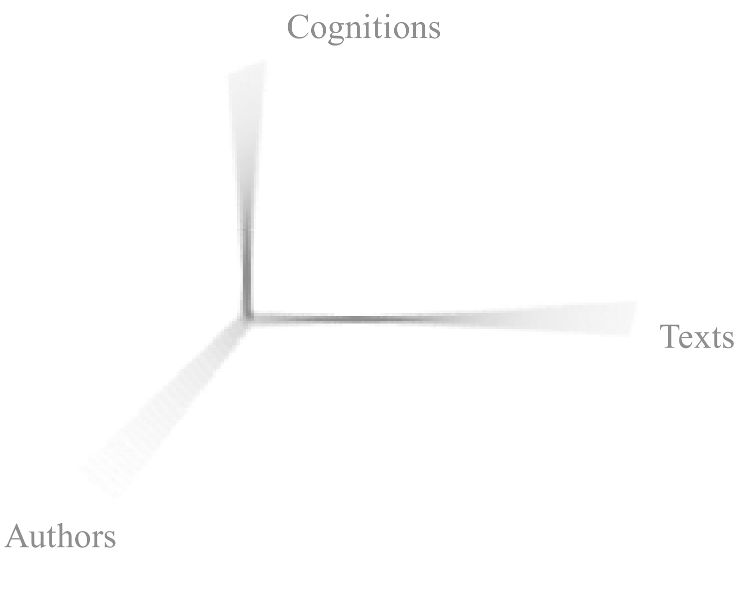
The Challenge of Scientometrics

The Development, Measurement, and Self-Organization of Scientific Communications



Loet Leydesdorff

The Challenge of Scientometrics:
The Development, Measurement, and Self-Organization
of Scientific Communications

Loet Leydesdorff. – uPUBLISH.COM: Universal Publishers.
Includes bibliographical references and index.
Subject headings:
self-organization, information sciences, communication theory, science studies, knowledge, scientometrics, indicators

First published by DSWO Press, Leiden University
Leiden, The Netherlands, 1995

Second edition.
Published by
Universal Publishers / uPUBLISH.com
USA • 2001

ISBN: 1-58112-681-6

http://www.upublish.com/books/leydesdorff-sci.htm

To learn more about this book visit the home page of the author at
http://www.leydesdorff.net/

TABLE OF CONTENTS

PART II
METHODOLOGICAL STUDIES USING INFORMATION THEORY

PREFACE

This study is the result of several years of collaboration as a scientometrician with philosophers, historians, and sociologists of science. It goes without saying that I am grateful to my colleagues in the Department of Science & Technology Dynamics of the University of Amsterdam for their relentless criticism and scepticism about scientometrics. In 1987, I had the opportunity to organize a workshop on the relations between qualitative theories in science and technology studies and the use of scientometric methods under the aegis of the European Association of Studies in Science and Technology (EASST), and to help edit a special issue of *Scientometrics* devoted to this subject (see: Leydesdorff *et al.* 1989). The program of study in this book is largely based on the research agenda that was formulated during this workshop.

Among the many colleagues, with whom I have discussed issues relevant to this study, I am particularly grateful to *Susan Cozzens* for several years of collaboration, and to *Michel Callon* and *Jean-Pierre Courtial* for discussions about the co-word methodology. In 1990, I spent some time as their guest at the Centre de Sociologie de l'Innovation of the École Nationale Supérieure des Mines in Paris. In the Dutch context, I wish to mention my colleagues *Gertrud Blauwhof*, *Peter Van den Besselaar* (Department of Social Science Informatics), *Wouter Van Rossum* (Groningen State University), and *Arie Rip* (Twente University) for discussions of scientometric methods and their theoretical interpretation. Finally, I am indebted to *Gene Moore* for correcting my English, but I remain responsible for any mistakes in the text.

Amsterdam, February 1995

Preface to the second edition

This second edition is not substantially different from the first one published by DSWO Press (Leiden University) in 1995. The text has been thoroughly revised, updated, and improved as necessary. I am grateful to Manfred Bonitz for spotting a number of errors and typos in the first edition. Among other things, I extended Chapter Ten with a new section about the implications of path-dependent transitions for firm behaviour and institutional agency (Blauwhof 1995; Leydesdorff and Van den Besselaar 1998).

The first edition has in the meantime been translated into Japanese by Yuko Fujigaki, Takayuki Hayashi, Hideyuki Hirakawa, Junichiro Makino, Masahi Shirabe, and Hiroyuki Tomizawa under the title *Saientometorikus no chôsen: kagaku-gijyutsu-joho no jiko-soshiki-ka* (Tokyo: Tamagawa University Press, 2001). The discussions with my Japanese colleagues were particularly intensive during the preparation of a special issue of *Scientometrics* on the 'Theory of Citations,' (Vol. 34, No. 1; see Leydesdorff 1998). In this context, I would also like to thank my colleague Paul Wouters for his contribution to what he has called *The Citation Culture* (Wouters 1999; Leydesdorff and Wouters 1999).

A further elaboration of my theory is available in *A Sociological Theory of Communications: The Self-Organization of the Knowledge-Based Society* (Universal Publishers, at http://www.upublish.com/books/leydesdorff.htm, 2001). The two books can be considered complementary in terms of providing theory and methods for the investigation of the knowledge base in processes of scientific communication and codification.

Loet Leydesdorff

Amsterdam,
February 2001

Chapter 1

Scientometrics and Science Studies

The tension between qualitative theorizing and quantitative methods is pervasive in the social sciences, and poses a constant challenge to empirical research. But in science studies as an interdisciplinary specialty, there are additional reasons why a more reflexive consciousness of the differences among the relevant disciplines is necessary.

First, the intellectual distance between contributions from the humanities, such as 'history of ideas' and philosophy, at one end of the spectrum of relevant disciplines, and from 'scientometrics' at the other end, is even more dramatic than in most social sciences; while an awareness of the differences in *methods* is particularly important because of the central position of the 'philosophy of science' in the constitution of the specialty. Second, in the past few decades, science studies has developed into an interdisciplinary specialty with its own journals, scholarly societies, and university departments. The consequent professional identity and ideology require a degree of integration of the insights from the various relevant disciplines, and the development of relatively independent and recognizable norms and standards in relation to neighboring disciplinary structures.

The span between disciplines which vary as much in terms of methods, standards and discursive styles as laboratory studies, intellectual history or scientometric indicators, is usually too large for the practice of empirical research. Within the framework of a single research project it will often prove inefficient or impractical to raise methodological questions concerning useful results from other disciplinary backgrounds. For example, historians who want to describe an intellectual lineage, and who may use patterns of citations or any other indicators to illustrate their arguments, are not usually interested in the possibility of clustering the same data with slightly different choices of methods into different structures that could shed further light on the object of study.

Decisions about provisional boundaries and methods are legitimate at the level of a project, or even at the level of an institutional program. However, intellectual exchanges at the level of the interdisciplinary specialty make a 'deconstruction' of the implied assumptions inevitable: what in one context appears as the practical assumptions of research may show a lack of sophistication and neglect of available knowledge when viewed from the perspective of another discipline. Without a common frame of reference, such discussions may easily disintegrate into priority disputes among participants from different programs and disciplines.

The commonality in the frame of reference in science studies has been formed mainly by a common interest in the subject matter, i.e., the development of the sciences. Theoretical integration has lagged behind because of the noted diversity among contributions from relevant disciplines. Efforts to integrate have taken the form of encyclopedic work, in which selections are made on the basis of pragmatic criteria, and collective efforts to produce handbooks, yearbooks, etc. (e.g., Spiegel-Rösing and De Solla Price 1977; Knorr-Cetina and Mulkay 1983; Van Raan 1988; Jasanoff *et al.* 1994).

Thus, it seems that interdisciplinary science studies are facing a dilemma. Theoretical justification is to be found in the various disciplinary backgrounds, while in debates within the specialty, these backgrounds can only function as legitimations for a particular approach. As soon as the approach is questioned, the discussion moves to a more philosophical level. But when one focuses on the results, the analyst seems to have no clear standards to evaluate them without provisionally accepting the approach. Capitalization on what the various contributions can teach us about the dynamics of science cannot be pursued systematically. This seems not a contingent choice: we actually lack methods for integration beyond the level of encyclopedic gathering. Indeed, the state of the art of science studies is 'pre-paradigmatic:' it is an interdisciplinary area integrated only at the level of its subject matter, and an applicational area for various contributing disciplines.

1.1 THE CHALLENGE OF SCIENTOMETRICS

As noted, the commonality in the intellectual enterprise of science studies is found in the commonality of the objects under study. Therefore, it seems appropriate to begin our search for a common framework at this end: what are the legitimate theoretical objects for science studies? What are the dimensions in which to phrase questions about them? How do we demarcate these questions from others which are not primary science-studies questions, although they may be of some relevance for the study of the sciences? These are epistemological questions concerning what should be considered as 'the world of science' as distinguished from other realms that can be studied in modern society.

The strength of the scientometrics program is its positive definition of science as an area of inquiry. The scientometric approach has often been reproached for its 'objective' pretensions (e.g., Edge 1979; Chubin and Restivo 1983). In my opinion, these pretensions are articulated with respect to particular methods and results, and one should not on this basis refute the challenge of scientometrics at the epistemological level, that is the claim that scientific developments are amenable to measurement.[1] I shall argue in this study that a multi-dimensional scheme like the one depicted in *Figure 1.1* can be used to describe this 'world of science.'

Along the three dimensions and their corresponding units of analysis, one can distinguish studies at various levels of

[1] The very possibility of defining science positively, and of making it subsequently an object of scientific investigations, is sometimes denied in the more reflexive tradition in science studies (e.g., Woolgar 1988).

aggregation.[2] For example, words are organized in texts, scientific articles in journals, journals belong to archives; scientists compose research groups, research groups belong to scientific communities; knowledge claims are based on theories, theories are embedded in disciplines. (One may wish to add more dimensions than the three indicated here.) The scheme suggests differences also in the nature of the dynamic processes along and between the axes (see, e.g., Holzner *et al.* 1987).

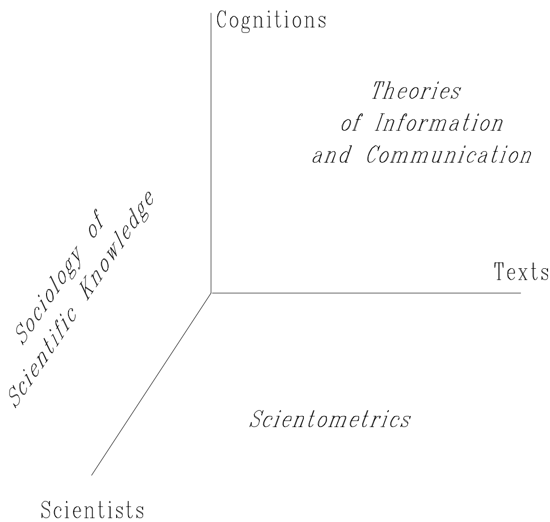


Figure 1.1
The study of the sciences as a multidimensional problem

In addition to a scheme which describes the types of objects and thus demarcates questions which we shall recognize as relevant to theorizing about the sciences, one is in need of a 'language' in which to study the phenomena within such a scheme. This 'language' should provide us with the methodological apparatus for describing this world coherently, despite the heterogeneity of the

[2] See for a similar categorization: Borgman 1989.

phenomena. Furthermore, the language should enable us to capture the core processes in scientific developments, and also guide us in the further choice of methods.

The pervasiveness of 'words' in science has previously led me and others to focus on words and co-words in such a comprehensive effort (Callon *et al.* 1986; Leydesdorff 1989b; Callon *et al.* 1993). In this study, I argue that 'information' is the more fundamental concept (cf. Mandelbrot 1968). The systematic processing of information in order to reduce uncertainty about the environment is the core process in scientific developments that the scientometrician attempts to map.

1.2 WORDS, CO-WORDS, INFORMATION, ENTROPY, SELF-ORGANIZATION

I proceed in two major steps: after a critical examination of qualitative and quantitative perspectives on science studies in Part One, a list of criteria can be composed for the methods which are needed for the development of science studies as an integrated enterprise. In Part Two, I shall show that information theory can comply with the listed criteria. By using this method, central problems in science studies will be addressed, both on the qualitative side (e.g., the significance of a reconstruction) and on the quantitative side (e.g., the prediction of science indicators).

'Information' (Shannon 1948) is yet content-free, which means that its content can still be defined at each level of aggregation, and in relation to the dimensions examined in a particular research design. Additionally, information as a measure is non-parametric, which means that we do not have to make any *a priori* assumptions concerning measurement scales or other mathematical idealizations (Krippendorff 1986). Furthermore, in its current formalization (Theil 1972), information theory is directly derived from probability theory, and, therefore, it is possible to relate results systematically to those of many other forms of social science statistics, and also to import results from the Bayesian philosophy of science. Finally, since all formulas in information theory are composed of simple summations, the use of these measures is highly appropriate with respect to the decomposition and/or aggregation.

The study of the sciences is so complex since the communication processes under study are multi-layered. Both the data and the (latent) structures in the data are in flux. Furthermore, the data can be considered as appreciations by scientists who are able to revise their interpretations reflexively. But changes in the data cannot be distinguished systematically from changes in the relevant dimensions–or more generally, changes at different levels–unless algorithmic methods of data analysis are used. Information calculus enables us to combine the multi-variate analysis of complex data structures (e.g., networks) with time-series analysis in a single design.

In Part Three, the study of possible irreversibilities in networks will lead me to the second major step in this study, namely to second-order systems theory, the theory of self-organization, and eventually to the specification of a mathematical sociology of scientific knowledge as a process of codification of scientific communication. It will be shown that the delineation of a complex unit of analysis from its contexts is a prerequisite for the prediction of the future behavior of the system(s) under study. This delineation of contexts remains necessarily hypothetical: structural developments are latent, and therefore, they can only be declared on the basis of an uncertain reconstruction. The analyst observes the interactions or 'instantiations' (Giddens 1979) of complex dynamic systems; the observations enable us to update our expectations.

1.3 ORGANIZATION OF THE STUDY

The scientometric conceptualization of science as a multi-dimensional construct that is susceptible to measurement is embedded in a philosophy of science. This philosophical position needs justification. The general organization of this study in different parts reflects this need of, on the one hand, justification and demarcation, and on the other, of methodological analysis and perspectives for empirical research.

Part One contains the theoretical justification of the program of the study. The multi-dimensional scheme is developed and assessed in relation to philosophical and sociological perspectives. In the next chapter, I analyze some major programs in the sociology of science, and show how methodological issues emerge as soon as sociologists do not limit their domain to the institutional dimensions of the scientific enterprise, but develop a sociology of scientific knowledge as well. I argue that in the sociology of scientific knowledge, important methodological problems have been reflected upon, but not yet been sufficiently clarified. Authors have coped with these problems by making strongly programmatic assumptions (e.g., Bloor 1976; Callon *et al.* 1983; Latour 1987a).

In my opinion, more rigorous distinctions among analytical dimensions, and between static and dynamic questions are needed. For example, the 'socio-cognitive' (inter-)action has become central in the new sociology of scientific knowledge.[3] The analyst, however, should disentangle the question of how the social and the cognitive dimensions co-vary in 'socio-cognitive' (inter-)action at any given moment in time, and the dynamic question of how action shapes and reproduces structure at a next moment. Both questions can be made subject to specification, and then be combined. The specific limitations introduced by programmatic assumptions with respect to the relations between these questions can also be specified.

One observes socio-cognitive interactions, but what these interactions mean can only be specified if one has hypothesized contexts in which these interactions can be provided with a meaning. The specification of the potentially different meanings of the interactions for a social context, a field of science, and/or other (e.g., subsequent) interactions, requires that one distinguish between a social variation, a cognitive variation, and a socio-cognitive co-variation or interaction term. The various effects cannot be expected to coincide, and thus, asymmetry prevails. Since sociologists of

[3] See, for example, Pinch (1982), at p. 17: "Within this interpretation 'paradigm' is taken to be a term which emphasizes the combined *socio-cognitive* nature of scientific activity."

scientific knowledge have programmatically argued in favor of symmetry in explaining the effects of 'socio-cognitive interaction,' this analytical conclusion may have far-reaching consequences.

Among other things, *Figure 1.1* (above) has provided us with a scheme that implies the declaration of an *analytically* independent cognitive dimension. This assumption, however, has consequences for research designs and the interpretation of results. How can a cognitive unit of analysis (e.g., a theory) be delineated in empirical research? In Chapter Three, I analyze some of the major traditions in modern philosophy of science by focusing on the methodological question of what is being explained about science in terms of what. On the one hand, the cognitive content of scientific knowledge is made central in philosophies of science that rooted in or react to logical positivism (notably, critical rationalism). In this context, the cognitive dimension is made the essential 'why' of everything else in the scientific enterprise. Cognitive developments function both as *explanans* for what is happening in science in all other dimensions, and as a normative criterion for distinguishing between what is in need of an explanation as a contribution to science and what is not.

For example, when Lakatos (1970) discussed the choices made by Niels Bohr when he developed the model for the atom as a research program, he explained Bohr's choices with hindsight in terms of what we know to have become the accepted model of the atom. The behavior of the scientist as an actor–in terms of choices with respect to lines of research–was explained in terms of the cognitive development of physics. However, the behavioral aspects are circumstantial in Lakatos' philosophy of science, and therefore, the philosopher's aim is not to explain these aspects, but to use the historical examples as only an illustration of the reconstruction in the cognitive dimension. Similarly, an illustration in the realm of the relations among scientific texts would have been possible, and notably equivalent in terms of its methodological status. Both the behavior of actors and the texts may serve as circumstantial evidence for ongoing theoretical developments.

In the other main tradition of the philosophy of science–the (neo-)conventionalist one–there is no such methodological equivalence between the content of texts and the behavior of actors.

They have a different status; here, language is the medium in which science develops, and therefore texts and discourse have a privileged position in the explanatory scheme. Science is, according to Quine (1962), a fabric of fact and theory, and correspondingly, any logical gap between theory and observation is reduced to matters of warp and weft, that is to matters of degree within language, and not of kind (Hesse 1980). The Popperian asymmetry between basic statements grounded on conventions versus theories located in World Three (or, analogously, Carnap's distinction between observational and theoretical statements) is now explicitly denied. In the neo-conventionalist tradition, one cannot even talk about explaining cognitions in terms of discourse, since the distinction between cognitions and language is now problematic. The relations among language and community are the remaining focus of interest.

The denial of the possibility of a separation between the cognitive and the linguistic dimension in the conventionalist tradition of the philosophy of science may seem attractive from the point of view of designing empirical research projects. The reduction of complexity, which is then possible, has been empirically fruitful: the question of what constitutes cognitive structure, i.e., its epistemological or even ontological status, can be neutralized as beyond the scope of empirical research, and therefore relatively irrelevant. Among others, adherents of the 'sociology of translation' or the actor-network approach have built heavily on these philosophical positions (e.g., Law and Lodge 1984).

I shall argue at two levels *against* a sociological reduction of the multi-dimensional problem to only the two dimensions of the literary manifestations of the sciences and the perceptions by local actors or groups. In Chapters Two and Three my argument is formulated at the theoretical level, and in Chapters Four and Five it will be shown empirically why one runs into problems if one focuses exclusively on observables like words, their co-occurrences or human actions. By using the full texts of eighteen scientific articles in a limited domain of biochemistry it can be shown that variation among word distributions is a result of various types of variation–conceptual variation, semantic variation, etc.–which have to be distinguished (Chapter Five). This distinction among independent sources of variation reintroduces the analysis of latent factors as a problem. In addition to their identification at each moment in time, one can raise questions about how the various factors change over time, in relation to change in the observable data. As we shall see, the declaration of change in latent factors requires the use of an algorithmic *calculus*.

The discussion of the various programs will make it possible (in Chapter Six) to list the requirements for a useful methodology of science studies. In addition to more technical requirements, methods should, for example, allow for the use of qualitative data and dynamic analysis, and not be restricted to contributions from specific theoretical perspectives in advance. I argue that there is scope for the development of such methods: probabilistic reasoning, because of its extensions in such a wide range of relevant disciplines as, e.g., information theory, statistical decomposition analysis, loglinear modelling, and Bayesian statistics, offers a perspective to develop a single comprehensive framework in which contributions from a great variety of disciplinary perspectives on science can be absorbed.

In Part Two of the study, I demonstrate the strength of using the relatively simple statistics of information theory to study some major problems of science studies. In Chapters Seven and Eight a static and a dynamic analysis of relations among the eighteen texts used in the study of word-distributions (from Chapter Five) is pursued, using information theory. However technical these studies may seem, their implications for empirical science and technology studies are substantive. As soon as the phenomena to be studied can be specified in empirical terms, the proposed methods can be applied to address issues such as (i) how much each unit (case or variable) accounts for the variation, (ii) the effects of aggregation and disaggregation, and (iii) in the dynamic model, questions concerning reconstructions.

The exploitation of these advantages will lead me from Chapter Nine onwards to reflections concerning the distinction between complexes of data which develop as systems, and those which do not. Chapter Ten focuses on how to study systems not in terms of relations but in terms of operations. Irreversibilities in network structures ('path-dependencies' and 'emergence') are discussed in probabilistic terms. Theoretically, this enables us to operationalize concepts in the actor-network approach (cf. Callon *et al.* 1986; Chapters Ten and Eleven), and to add the time-dimension to the 'structural theory of action' (cf. Burt 1982; Chapter Twelve).

In Part Three, the notion of developing systems is addressed more explicitly. In Chapter Eleven, the impact of EC science policies on the transnational publication system in Western Europe is analyzed empirically: did a European system emerge in addition to the various national research systems? This research question reintroduces the multi-variate perspective in the dynamic analysis: if an actor-network is not one system, but a composite of separate systems (actors) in a network with potentially different operations, how then are we able to study the interactions between these (e.g., national) systems?

The assumption that the actors at the nodes can operate with relative independence from the operation of the network is a familiar model in parallel and distributed computing: each processor performs its own operations while the network runs a different program (e.g., Rumelhart *et al.* 1986; Bertsekas and Tsitsiklis 1989). This model is also used in artificial intelligence as a solution to the problem of a local update in the light of specific information (Pearl 1988). The methodological toolbox which we have created in Part Two provides us with an operationalization of a notion of structure as contingent, but in its distributed operation only dependent upon other contingencies insofar as the latter operate. Otherwise, self-referentiality (operationalizable in terms of auto-covariation) prevails. In Chapter Twelve, this program for empirical science studies is delineated from the Bayesian program in the philosophy of science, and from the use of knowledge representations in artificial intelligence.

Luhmann (1984) elaborated a model for society as a communication system. The social system is no longer understood as an aggregate of human beings, but as the system of links which is added to and contingent upon the nodes (i.e., individuals) that perform their own operations. Luhmann's (1990) sociology of science and Shannon's (1948) mathematical theory of communication share a common background in modern biology and non-equilibrium thermodynamics (e.g., Prigogine and Stengers 1979/1984; Maturana and Varela 1980; cf. Swenson 1989). However, they are not just another application of the principles of thermodynamics: they reveal with hindsight that the study of the sciences at the meta-level is itself part of the development of the sciences, and therefore can be linked with current developments in methodologies, i.e. the study of complex systems that are not in equilibrium. In a final chapter, entitled 'The Possibility of a Mathematical Sociology of Scientific Communications,' I specify these conclusions of the study in relation to other traditions in science studies.[4]

---

[4] See for the elaboration at the level of the social system: *A Sociological Theory of Communication: The Self-Organization of the Knowledge-Based Society.* Universal Publishers, at http://www.upublish.com/books/leydesdorff.htm, 2001.

PART I

THEORETICAL PERSPECTIVES ON SCIENTOMETRICS

Chapter 2

Scientometrics and the Sociology of Science

Until around 1970, questions about the growth and dynamics of scientific knowledge belonged to the realm of philosophy. The central issue in the philosophy of science was the *validity* of knowledge (the 'context of justification'). The philosophical reconstruction, however, was analytically to be distinguished from questions about how that knowledge was being produced (the 'context of discovery'). The latter realm was believed to belong to the domain of the social sciences.

The link between the philosophical issue of the growth of scientific knowledge and the sociological quest for explanations of variance in observable distributions was in large part established by historians like *Price* and *Kuhn*, who were able to see the substantive developments in the wider contextual perspective of the institutional growth of the scientific enterprise. Price (1965) emphasized the relations between knowledge growth and document sets; Kuhn (1962) highlighted the relations between authors working within paradigms and the growth of knowledge.

Studies concerning the relations between document sets and groups of authors constitute a natural extension of the set of questions accessible to the multi-dimensional scheme which was introduced in the previous chapter (e.g., Crane 1969), although these questions may be less obvious from the perspective of writing the intellectual history of science. This extension, however, provided a bridge between bibliometric approaches and sociological theorizing in science studies (e.g., Griffith and Mullins 1972; Cole and Cole 1973). In particular, following the proposal by Small and