Springer Complexity

Springer Complexity is an interdisciplinary program publishing the best research and academic-level teaching on both fundamental and applied aspects of complex systems — cutting across all traditional disciplines of the natural and life sciences, engineering, economics, medicine, neuroscience, social and computer science.

Complex Systems are systems that comprise many interacting parts with the ability to generate a new quality of macroscopic collective behavior the manifestations of which are the spontaneous formation of distinctive temporal, spatial or functional structures. Models of such systems can be successfully mapped onto quite diverse "real-life" situations like the climate, the coherent emission of light from lasers, chemical reaction-diffusion systems, biological cellular networks, the dynamics of stock markets and of the internet, earthquake statistics and prediction, freeway traffic, the human brain, or the formation of opinions in social systems, to name just some of the popular applications.

Although their scope and methodologies overlap somewhat, one can distinguish the following main concepts and tools: self-organization, nonlinear dynamics, synergetics, turbulence, dynamical systems, catastrophes, instabilities, stochastic processes, chaos, graphs and networks, cellular automata, adaptive systems, genetic algorithms and computational intelligence.

The three major book publication platforms of the Springer Complexity program are the monograph series "Understanding Complex Systems" focusing on the various applications of complexity, the "Springer Series in Synergetics", which is devoted to the quantitative theoretical and methodological foundations, and the "SpringerBriefs in Complexity" which are concise and topical working reports, case-studies, surveys, essays and lecture notes of relevance to the field.

In addition to the books in these two core series, the program also incorporates individual titles ranging from textbooks to major reference works.

Editorial and Programme Advisory Board

Understanding Complex Systems

Founding Editor: J.A. Scott Kelso

Future scientific and technological developments in many fields will necessarily depend upon coming to grips with complex systems. Such systems are complex in both their composition – typically many different kinds of components interacting simultaneously and nonlinearly with each other and their environments on multiple levels – and in the rich diversity of behavior of which they are capable.

The Springer Series in Understanding Complex Systems series (UCS) promotes new strategies and paradigms for understanding and realizing applications of complex systems research in a wide variety of fields and endeavors. UCS is explicitly transdisciplinary. It has three main goals: First, to elaborate the concepts, methods and tools of complex systems at all levels of description and in all scientific fields, especially newly emerging areas within the life, social, behavioral, economic, neuro- and cognitive sciences (and derivatives thereof); second, to encourage novel applications of these ideas in various fields of engineering and computation such as robotics, nano-technology and informatics; third, to provide a single forum within which commonalities and differences in the workings of complex systems may be discerned, hence leading to deeper insight and understanding.

UCS will publish monographs, lecture notes and selected edited contributions aimed at communicating new findings to a large multidisciplinary audience.

For further volumes:
http://www.springer.com/series/712

Andrea Scharnhorst • Katy Börner
Peter van den Besselaar

Models of Science Dynamics

Encounters Between Complexity Theory and Information Sciences

Springer

Editors
Dr. Andrea Scharnhorst
The Virtual Knowledge Studio
for the Humanities and Social Sciences
Royal Neth. Academy of Arts a.
Sciences
Cruquiusweg 31
1019 AT Amsterdam
Netherlands
andrea.scharnhorst@vks.knaw.nl

Katy Börner
Cyberinfrastructure for Network
Science Center
School of Library and Information Sc.
Indiana University
10th Street and Jordan Avenue
47405 Bloomington
USA
katy@indiana.edu

Peter van den Besselaar
Science System Assessment Center
The Rathenau Institute
Anna van Saksenlaan 51
2593 HW The Hague
Netherlands

ISSN 1860-0832 e-ISSN 1860-0840
ISBN 978-3-642-23067-7 e-ISBN 978-3-642-23068-4
DOI 10.1007/978-3-642-23068-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: "PCN Applied for"

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To advance the future of
science of science models

Foreword

Andrea Scharnhorst, Data Archives and Networked Services & e-Humanities group, The Royal Netherlands Academy of Arts and Sciences, The Netherlands

Katy Börner, Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, USA

Peter van den Besselaar, Department of Organization Science & Network Institute, VU University Amsterdam, The Netherlands

Motivation

Models of Science Dynamics aims to capture the structure and evolution of science – scholars and science itself become "research objects." These research objects might be represented by conceptual models based on historical and ethnographic observations, mathematical descriptions of measurable phenomena, or computational algorithms. Some models re-create the structure of co-authorship networks and their evolution over time. Others capture the dynamics of citation diffusion patterns.

The philosophy, history, and sociology of science have produced valuable insights into the nature of scholarly activities as a human activity and social system. Within this area, the dynamics and structure of the science system, including the social sciences and humanities, have been the focus of a variety of explanatory, exploratory, and metaphorical models (Kuhn 1962; Cole and Cole 1967; Crane 1972; Elkana 1978; Nowakowska 1984; Price 1963; Nalimov and Mulchenko 1969; Leydesdorff and Van den Besselaar 1997). Almost every progress in mathematical modeling has also been applied to model science itself. Phenomena such as specific growth laws of publications and citations (Price 1965, 1976), scientific productivity (Lotka 1926), or the distribution of topics over journals (Bradford 1934) have always raised the interest of mathematicians and natural scientists. Mathematical models have been proposed not only to explain statistical regularities (Egghe and Rousseau 1990), but also to model the spreading of ideas (Goffman 1966) and the competition between scientific paradigms (Sterman 1985) and fields (Kochen

1983; Yablonskiĭ 1986; Bruckner et al. 1990). Furthermore, they have been used to model the relation between publishing, referencing, and the emergence of new topics (Gilbert 1997), as well as the co-evolution of co-author and paper-citation networks (Börner et al. 2004; Börner and Scharnhorst 2009; Börner 2010). Outside of the field of science and technology studies, such models have also been presented and discussed at conferences about self-organization, system dynamics, agent-based modeling, artificial societies, and complexity theory. Despite its evident importance, however, the mathematical modeling of science still lacks a unifying framework and a comprehensive study of the topic. This book aims to fill this gap.

Structure of the Book

This book reviews and describes major threads in the mathematical modeling of science dynamics for a wider academic audience. The model classes presented cover stochastic and statistical models, system-dynamics approaches, agent-based simulations, population-dynamics models, and complex-network models. The book starts with an introduction (Börner et al. 2011) and a foundational chapter that defines and operationalizes terminology used in the study of science. This is followed by a review chapter (Lucio-Arias and Scharnhorst 2011) that discusses the history of mathematical approaches to modeling science from an algorithmic-historiography perspective. The subsequent chapters review specific modeling approaches such as population-dynamic (Vitanov and Ausloos 2011), agent-based (Payette 2011), and game-theoretic models (Hanauske 2011). Different modeling approaches used to capture the structure and dynamics of social networks (Mali et al. 2011) and citation networks (Radicchi et al. 2011) are presented in two separate chapters. Model classes often combine descriptive and predictive elements—this book places a strong emphasis on the latter. The book concludes with a short outlook (van den Besselaar et al. 2011) to remaining challenges for future science models and their relevance for science and science policy.

References

Börner K (2010) Atlas of science: Visualizing what we know. MIT, Cambridge, MA

Börner K, Boyack KW, Milojević S, Morris S (2011) Basic model types, key definitions, and a general framework for the comparison of process models. In: Scharnhorst A, Börner K, AQ1 van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 1, pp xx–xx

Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. Proc Natl Acad Sci USA 101(Suppl 1):5266–5273 (DOI: 10.1073/pnas.0307625100), also available as arXiv preprint arXiv:cond-mat/0311459. The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci USA 101(Suppl 1):5183–5310

Börner K, Scharnhorst A (2009) Visual conceptualizations and models of science. Editorial for the special issue ‘Science of science: Conceptualizations and models of science’.

J Informetr 3(3):161–172 (DOI: 10.1016/j.joi.2009.03.008), also available as arXiv preprint arXiv:0903.3562

Bradford SC (1934) Sources of information on specific subjects. Eng Illustrated Weekly J 137(3550):85–86. Reprinted: (1985) J Inform Sci 10(4):176–180 (DOI: 10.1177/016555158501000406)

Bruckner E, Ebeling W, Scharnhorst A (1990) The application of evolution models in scientometrics. Scientometrics 18(1):21–41 (DOI: 10.1007/BF02019160)

Cole S, Cole JR (1967) Scientific output and recognition: A study in the operation of the reward system in science. Am Sociol Rev 32(3):377–390 (DOI: 10.2307/2091085, stable JSTOR URL: http://www.jstor.org/stable/2091085)

Crane D (1972) Invisible colleges: Diffusion of knowledge in scientific communities. University of Chicago Press, Chicago, IL

Egghe L, Rousseau R (1990) Introduction to informetrics: Quantitative methods in library, documentation and information science. Elsevier, Amsterdam

Elkana Y, Lederberg J, Merton RK, Thackray A (eds) (1978) Toward a metric of science: The advent of science indicators. Science, Culture and Society. Wiley, New York, NY

Gilbert N (1997) A simulation of the structure of academic science. Sociol Res Online 2(2), available online at the URL: http://www.socresonline.org.uk/2/2/3.html

Goffman W (1966) Mathematical approach to the spread of scientific ideas – the history of mast cell research. Nature 212(5061):449–452 (DOI: 10.1038/212449a0)

AQ2 Hanauske M (2011) Evolutionary game theory and complex networks of scientific information. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 5, pp xx–xx

Kochen M (1983) Mathematical model for the growth of two specialties. Science of Science 3(11):199–217

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago, IL

Leydesdorff L, van den Besselaar P (1997) Scientometrics and communication theory: Towards theoretically informed indicators. Scientometrics 38(1):155–174 (DOI: 10.1007/BF02461129), also available as arXiv preprint arXiv:1001.4433

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12): 317–323

AQ3 Lucio-Arias D, Scharnhorst A (2011) Mathematical approaches to modeling science from an algorithmic-historiography perspective. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 2, pp xx–xx

AQ4 Mali F, Kronegger L, Doreian P, Ferligoj A (2011) Dynamic scientific co-authorship networks. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 6, pp xx–xx

Nalimov VV, Mul'chenko ZM (1969) Naukometriya. Nauka, Moscow

Nowakowska M (1984) Theories of research, vols I–II. Systems Inquiry Series. Intersystems Publications, Seaside, CA

AQ5 Payette N (2011) Agent-based models of science. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 4, pp xx–xx

Price DJ de Solla (1963) Little science, big science. Columbia University Press, New York, NY

Price DJ de Solla (1965) Networks of scientific papers. Science 149(3683):510–515 (DOI: 10.1126/science.149.3683.510). Reprinted in: Price DJ de Solla (1986) Little science, big science . . . and beyond. Columbia University Press, New York, NY, pp 103–118

Price DJ de Solla (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306 (DOI: 10.1002/asi.4630270505), also available online at the URL: http://www.asis.org/Publications/JASIS/Best_Jasist/1976pricejasistarticle.pdf

AQ6 Radicchi F, Fortunato S, Vespignani A (2011) Citation networks. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 7, pp xx–xx

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. Technol Forecast Soc Change 28(2):93–122 (DOI: 10.1016/0040-1625(85)90009-5) van den Besselaar P, Börner K, Scharnhorst A (2011) Science policy and the challenges for modeling science. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 8, pp xx–xx AQ7

Vitanov NK, Ausloos MR (2011) Knowledge epidemics and population dynamics models for describing idea diffusion. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 3, pp xx–xx AQ8

Yablonskiĭ AI (1986) Matematicheskie modeli v issledovanii nauki. Nauka, Moscow

AUTHOR QUERIES

AQ1. Please update ref. Börner et al. (2011)
AQ2. Please update ref. Hanauske (2011)
AQ3. Please udpate ref. Lucio-Arias and Scharnhorst (2011)
AQ4. Please update ref. Mali et al. (2011)
AQ5. Please update ref. Payette (2011)
AQ6. Please update ref. Radicchi et al. (2011)
AQ7. Please update ref. van den Besselaar et al. (2011)
AQ8. Please update ref. Vitanov and Ausloos (2011)

Preface

Andrea Scharnhorst, Data Archives and Networked Services & e-Humanities group, The Royal Netherlands Academy of Arts and Sciences, The Netherlands

Models of science – scattered knowledge

After World War II, scientists were increasingly subject to systematic and large scale measurements efforts. The growth and changing roles of science stimulated the need for governmental and "policy support of science" as well as the need for an empirical basis for "science policy". Since then a wealth of monitoring and evaluative indicators has been created. Sociology of science (Bernal 1939; Kuhn 1962; Merton 1973) as well as Scientometrics (Nalimov and Mulchenko 1969; Price 1963) were established as scientific fields. The Society for Social Studies of Science (4S), the European Associations for the Study of Science and Technology (EASST) and the International Society of Scientometrics and Informetrics (ISSI), among others, are active as professional organisations in this field. At their conferences "models of science" occasionally appear, but are not presented in a systematic way on a regular basis. Not only other knowledge domains, such as sociology, philosophy, economics, but also physics apply their models to science (Lucio-Arias and Scharnhorst 2011, Chap. 2), but so far there has been no common reference point such as a conference series, edited books, or monographs devoted to modeling science. The only exception to our knowledge, beyond review sections in journal articles (e.g., Börner et al. 2004), a review in ARIST (Börner et al. 2003), and a recent special issue on Science of Science (Börner and Scharnhorst 2009), is the monograph of Yablonskiĭ published 1986 in Russian with Nauka and not translated into English (Yablonskiĭ 1986 Matematicheskie Modeli v Issledovanii Nauki (in Russian) Nauka, Moscow (Yablonskiĭ 1986)). This edited volume aims to fill this gap by presenting an overview about major current trends in modeling of science (Chaps. 3 (Vitanov and Ausloos 2011), 4 (Payette 2011), 5 (Hanauske 2011), 6 (Mali et al. 2011), and 7 (Radicchi et al. 2011)) and a general framework to relate these trends to each other (Börner et al. 2011, Chap. 1).

New possibilities and challenges from information science – mapping science

This book is also an expression of a growing interest in the field of modeling science. One origin of this development can be found in recent achievements in information and computer sciences. They have made it possible to visualize research activities at an unprecedented scale and with a high level of sophistication (Börner et al. 2003). Networks of publications and their citation patterns, word use, collaborating researchers, or topics in e-mail threads have been measured, analysed and visualized over time. With the emergence of network science (Chaps. 6 (Mali et al. 2011) and 7 (Radicchi et al. 2011)) as a new cross-disciplinary approach (Barabási 2002; Barabási et al. 2002) and in particular with the achievements of visualizing knowledge domains in the information sciences (Shiffrin and Börner 2004), old dreams of mapping the sciences (Garfield et al. 1964; Small and Griffith 1974) can now be realized. A prominent example of this approach are the so-called "maps of science" which show all scientific disciplines—as far as their activities are covered by the ISI Thompson Reuters" Web of Knowledge, Elsevier's Scopus, or other databases (Boyack et al. 2005). A prominent initiative for mapping science is the NSF funded "Mapping Science" exhibit (http://scimaps.org) informing a wide audience about a new "cartography of science" (Börner 2010). The new maps of science inspire new models as explanatory tools for emergent structures of the science system. Mathematical models of complex systems play a specific role in this discourse.

Beyond mapping – towards explanations

Information gathering about science as a backbone of the knowledge society is only one aspect of these new developments. These instruments are also meant as tools to detect and maybe forecast conditions under which scientific discoveries emerge and areas where these discoveries can be found. At the same time, basic questions about the understanding of science are raised, such as who are the actors driving the development of science: individuals, groups or institutions. Earlier large-scale maps concentrated on scientific communications as manifested in papers and their citation interlinkage (Scharnhorst and Garfield 2010). Partly, this was due to the fact that unique author names are hard to determine because of same names, name variants and misspellings. So, a large part of bibliometrics and scientometrics analyses texts (titles, keywords, words, references). Some automated techniques have partly solved this problem, at least on a higher level of aggregation. In maps of scientific communication, authors as well as institutions can now be made visible with a higher reliability. To explain the networks in which researchers are linked (by publishing or communicating), current research in social-psychology and sociology of science becomes relevant. Resumé analysis, ethnographic observations, and interviews were presented as ways to gain access to local motivations and behavior. The collective effect of which is reflected in the large scale global maps of science. We call this the **return of the actors** in scientometrics research. If one thinks in terms of modeling network of scholars, these models entail assumptions about the behavior of the "nodes" (Mali et al. 2011, Chap. 6). This is the moment when qualitative, quantitative, and mathematical models need to come together.

A second observation concerns the increasing need to explain changes in scholarly activities. The design of mostly static maps of science, social science and the humanities is therefore only a starting point. Ultimately, we need to see and understand the dynamics of science (Börner et al. 2004; Leydesdorff and Schank 2008; Börner and Scharnhorst 2009). Visualizations that show the unfolding of scholarly activities in a 'fast forward' mode can help refute or confirm existing theories and trigger questions for novel research into the basic mechanisms of scientific growth. We call this *the return of time and dynamics.*

Contribution of models – models as heuristic devices. Meeting between information science and physics

Mathematical models represent a very specific *instrumentarium* to analyse elementary processes behind measurable phenomena on a more global scale. As mentioned above, in particular during the 1970s and 1980s, the science system has been conceptualized as a self-organizing system in sociology (Luhmann 1990) as well as modeled using concepts and techniques from physics and cybernetics (Scharnhorst 1988). Nowadays, network models are proposed for studying scientific collaborations or the emergence of topics. These new approaches to the modeling of science look into the growth of scholarly networks (Barabási et al. 2002; Committee on Network Science for Future Army Applications 2005; Börner et al. 2004), the structure of scientific communities (Newman et al. 2006), the epidemics of ideas on collaboration networks (Bettencourt et al. 2006), scholarly information foraging (Sandstrom 1994), the formation of effective teams (Amaral and Uzzi 2007), the competition of groups about paradigms (Chen et al. 2009), the scientific productivity of generations of scientists over time (Fronczak et al. 2007), and modeling the dynamics of actor networks (Snijders et al. 2007). However, as mentioned above, the many existing models of science have been developed in many different scientific fields ranging from physics, sociology to history of science. They exist often unrelated and independently from each other and are seldom linked to other studies of science. Nevertheless, in the last couple of years we have witnessed several encounters between physics and information sciences (Fortunato 2010; Bollen et al. 2009; Barabási 2002). This book aims to contribute to a consolidation of the knowledge about models and their mutual dependencies.

Outline of the book

The book consists of four parts: Part I – Foundations, Part II – Exemplary Model Types, Part III – Exemplary Model Applications and Part IV – Outlook.

Part I contains two chapters. In Chap. 1 "An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models" (Katy Börner, Kevin W. Boyack, Staša Milojević and Steven Morris). Börner et al. (2011) develop a set of reference or frames along which models can be ordered and compared. Departing from a general definition of the term "model" the authors identify a set of dichotomies, such as descriptive versus process models, which can be used to differentiate between essence, purpose and insights of different models. Even if the reader might want to extend or alter the prosed criteria, he or she has to accept that no comparison of models is possible without a clear articulation of their main elements (units, interactions, targeted phenomena) and their tentative ordering in a common reference framework. With a glossary at the end of his chapter, the authors further deliver jigsaw pieces for a common ground on which models can be related to each other.

One cannot understand the emergence and the essence of certain models without looking into the history of modeling science. The emphasis of certain perspectives of modeling science above others is obviously correlated with the overall *Zeitgeist* of a certain time period. Accordingly, the second chapter ("Mathematical approaches to modeling science from an algorithmic-historiography perspective " by Diana Lucio-Arias and Andrea Scharnhorst) (Lucio-Arias and Scharnhorst 2011) describes the history of science models combining a participant story with a bibliometric reconstruction. Histories are always told on the basis of a set of experiences on the one side and a set of norms and values on the other. Consequently, a variety of histories can be found. Only recently the different perception of members of a scientific community could be made visible by a bibliometric analysis of the citation network of this community (Havemann et al. 2010). Chapter 2 (Lucio-Arias and Scharnhorst 2011) chooses the classical method of algorithmic historiography as introduced by Eugene Garfield. One of the most interesting findings is that current threads in mathematical modeling in scientometrics seem to ignore each other while at the same time relying on the same classical papers.

Part II – Exemplary Model Types contains three chapters which all review models belonging to a certain class of mathematics and partly also introduce own model approaches. We are quite aware that these chapters do not cover all occurring threads in the history and presence of science models. Missing are, for example, system dynamics (Sterman 1985) which has been successfully applied in innovation studies and urban development, or entropy and information measures. The threads reviewed in this part of the book are examples in which selection is based on the availability of authors, and of course which could be extended. Although they all use an individual language, what binds them together is a more generic perspective of science models. All chapters depart from mathematical techniques available and interrogate to which extent they can be used to obtain a better understanding of the science system. Accordingly, the empirical validation of the models is discussed but not in the foreground. These chapters introduce the reader to the details of the model building process in terms of conceptualization, abstraction, operationalization and extension towards increasingly more complex models. In Chap. 3 (Knowledge epidemics and population dynamics models for describing idea diffusion) Nicolai Vitanov and Marcel Ausloos (Vitanov and Ausloos 2011) present a rich inventory of dynamic models based on the behavior of groups of scientists and suitable to describe the emergence and spreading of new ideas in a competitive process. Groups of scientists can be defined based on their actual acquaintance with a certain idea (epidemic models) or their membership in a certain scientific community. That scientists can change their membership in scientific communities creates an extra challenge for modeling. The authors also discuss the role of fluctuations during the emergence of innovation and when best to turn from deterministic models to more complex stochastic models. This chapter also demonstrates that a further methodological exploration is needed to fill the toolbox of science models. With this respect in mind the introduction of time-lag elements and the combination of time and space are the most original contributions in this chapter. Nicolas Payette (2011) introduces the reader in Chap. 4 "Agent-Based Models of Science" into the world of agent-based modeling as practiced in computational sociology and computational philosophy. Obviously, the type of rule based modeling as proposed by Epstein and Axtell connects very well to known social theories about the behavior of social beings. Payette digs out the longer history of agent-based modeling, which goes back to John von Neumann. Actually, there are links to spin models (widely applied in sociophysics) waiting for further exploration (Stauffer and Solomon 2007). The chapter provides the reader with excellent and clear insights into the inner logic of different ABM approaches to science. In difference to dominant mathematical language of the previous chapter, in an interesting contrast, Payette compares models qualitatively by mapping their different conceptual frames. He highlights possible links to other model threads such as network models. Matthias Hanauske returns in Chap. 5 "Evolutionary Game Theory and Complex Networks of Scientific Information" (Hanauske 2011) to the power of mathematics and scientific diagrams. Triggered by a real-world phenomenon – the reorganization of the market of scientific publishing – Hanauske questions the possibilities to model the interaction of different players in this process (authors and scientific journals) with game theory. Game theory is designed to explore the consequence of individual strategic behavior in interactions between many individuals. In particular it allows statements for multi-level networked systems – a suitable description for the complex interaction of producers and disseminators of scientific products where the same individuals often switch roles.

Part III – Exemplary Model Applications describes models for two major aspects of scientific communication: co-authoring and referencing. Not surprisingly a network model approach is applied to both phenomena, relying on the different epistemic traditions of sociology and physics. Co-authoring and referencing are both part of scientific production. Consequently, in Chap. 6 “Dynamic Scientific Co-Authorship Networks”, Mali et al. (2011) start with the whole universe of scientific communication before zooming into their specific topic of co-authoring. They also start with an excellent history of Social Network Analysis. Here the reader is provided with detailed context to obtain a better understanding of the sources of some of the still existing tensions between different network approaches. Among the dynamic models, blockmodeling applied to evolving networks and stochastic actor-oriented models form the cornerstones of this chapter. Empirical studies are extensively reviewed; ordered alongside of dimensions of cross-disciplinary, cross-sectoral and cross-national collaboration pattern; and linked to SNA model insights. Among their own studies one of the interesting findings points to a tension between strongly local (national) connectivity and the requirements of being interwoven into the international (global) knowledge production. Chapter 7, “Citation Networks” of Radicchi et al. (2011) complementary to Chap. 6 (Mali et al. 2011) looks into (citation) networks from a statistical physics perspective. Again we see a recurrent pattern. Following the epistemic tradition of physics, Radicchi et al. (2011) insist on the search for universality and general organizing principles in their network studies where Mali et al. (2011), in the epistemic tradition of sociology, emphasize how best to incorporate the multi-facet roles of individuals in networks and the different context of their link structures. Nevertheless, there is an overlapping area. Against expectations based on the knowledge of how different the citation behavior is in different disciplines, on a statistical level there are still similarities or universalities. It remains open if these ‘general laws’ are just mathematical artefacts or if the point to a shared feature in citing across disciplines. Also in SNA the aim is to detect a general pattern in social behavior (as for instance by blockmodeling). In both cases the challenge is to give these patterns a meaningful interpretation. Similar to (Mali et al. 2011, Chap. 6), the authors of Chap. 7 (Radicchi et al. 2011) carefully discuss empirical material. Time is a leading theme through both chapters. Time is the ‘hidden constructor’ behind specific distributions of networks (such as degree distributions). The authors of Chap. 7 address time more explicitly in dynamic models of the evolution of citation networks and diffusion processes across citation networks. Concerning the latter they take a very elegant and original approach – namely to model papers in terms of their received reward by citations. While citation networks are cumulative in time and the position of a paper in such a network cannot change, its perception can change with each new generation of citing papers; so, reward and recognition of a paper can travel in network topologies and in this way, the diffusion of ideas become visible.

The book concludes with Part IV – Outlook. Chapter 8 “Science policy and the challenges for modeling science” partly also reflects on the process of the making of the book, and the lessons learned from it (van den Besselaar et al. 2011). Despite the character of the book as a collection of chapters, authors and editors have taken specific measures to enhance the consistency of it. This becomes visible in the different appendices of the book. A glossary of relevant terms comes as appendix with Börner et al. (2011) Chap 1. Another group of appendices lists the (historic) knowledge base of the field – adding details to Lucio-Arias and Scharnhorst 2011, Chap. 2. Also all model chapters in Parts II and III contain overviews and short descriptions of the models they address. They also contain text boxes (Key points) highlighting main insights for the general audience and/or science policy makers.

References

Amaral LAN, Uzzi B (2007) Complex systems – a new paradigm for the integrative study of management, physical, and technological systems. Manag Sci 53(7):1033–1035 (DOI: 10.1287/mnsc.1060.0696)

Barabási AL (2002) Linked: The new science of networks. Perseus, Cambridge, MA

Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311(3–4):590–614 (DOI: 10.1016/S0378-4371(02)00736-7), also available as arXiv preprint arXiv:cond-mat/0104162

Bernal JD (1939) The social function of science. George Routledge & Sons, London

Bettencourt LMA, Cintrón-Arias A, Kaiser DI, Castillo-Chávez C (2006) The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A 364:513–536 (DOI: 10.1016/j.physa.2005.08.083), also available as arXiv preprint arXiv:physics/050206

Bollen J, van de Sompel H, Hagberg A, Bettencourt L, Chute R, Rodriguez MA, Balakireva L (2009) Clickstream data yields high-resolution maps of science. PLoS ONE 4(3):e4803, 11 pp (DOI: 10.1371/journal.pone.0004803)

Börner K (2010) Atlas of science: Visualizing what we know. MIT Press, Cambridge, MA

Börner K, Boyack KW, Milojević S, Morris S (2011) Basic model types, key definitions, and a general framework for the comparison of process models. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 1, pp xx–xx

Börner K, Chen C, Boyack KW (2003) Visualizing knowledge domains. Ann Rev Inf Sci Technol 37:179–255 (DOI: 10.1002/aris.1440370106)

Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. Proc Natl Acad Sci USA 101(Suppl 1):5266–5273 (DOI: 10.1073/pnas.0307625100), also available as arXiv preprint arXiv:cond-mat/0311459. The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci USA 101(Suppl 1):5183–5310

Börner K, Scharnhorst A (2009) Visual conceptualizations and models of science. Editorial for the special issue 'Science of science: Conceptualizations and models of science'. J Informetr 3(3):161–172 (DOI: 10.1016/j.joi.2009.03.008), also available as arXiv preprint arXiv:0903.3562

Boyack KW, Klavans R, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3):351–374 (DOI: 10.1007/s11192-005-0255-6)

Chen C, Chen Y, Horowitz M, Hou H, Liu Z, Pellegrino D (2009) Towards an explanatory and computational theory of scientific discovery. J Informetr 3(3):191–209. (DOI: 10.1016/j.joi.2009.03.004), also available as arXiv preprint arXiv:0904.1439

Committee on Network Science for Future Army Applications NRC (2005) Network science. The National Academies Press, Washington, DC, available online at the URL: http://www.nap.edu/catalog.php?record_id=11516

Fortunato S (2010) Community detection in graphs. Phys Reports 486(3–5):75–174 (DOI: 10.1016/j.physrep.2009.11.002), also available as arXiv preprint arXiv:0906.0612

Fronczak P, Fronczak A, Hołyst JA (2007) Publish or perish: Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. Phys Rev E 75(2):026103, 9 pp (DOI: 10.1103/PhysRevE.75.026103), also available as arXiv preprint arXiv:physics/0606190

Garfield E, Sher IH, Torpie RJ (1964) The use of citation data in writing the history of science. Technical report, Institute for Scientific Information Inc. Philadelphia, PA. Report of research for the Air Force Office of Scientific Research under contract AF49 (638)-1256, also available online at the URL: http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf

AQ1 Hanauske M (2011) Evolutionary game theory and complex networks of scientific information. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 5, pp xx–xx

Havemann F, Heinz M, Struck A, Gläser J (2010) Identification of overlapping communities by locally calculating community-changing resolution levels. Arxiv preprint arXiv:1008.1004

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago, IL

Leydesdorff L, Schank T (2008) Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. J Am Soc Inf Sci Technol 59(11):1532–2890 (DOI: 10.1002/asi.20891), also available as arXiv preprint arXiv:0911.1437

AQ2 Lucio-Arias D, Scharnhorst A (2011) Mathematical approaches to modeling science from an algorithmic-historiography perspective. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 2, pp xx–xx

Luhmann N (1990) Die Wissenschaft der Gesellschaft. Suhrkamp, Frankfurt am Main

AQ3 Mali F, Kronegger L, Doreian P, Ferligoj A (2011) Dynamic scientific co-authorship networks. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 6, pp xx–xx

Merton RK (1973) The sociology of science: Theoretical and empirical investigations. University of Chicago Press, Chicago, IL

Nalimov VV, Mul'chenko ZM (1969) Naukometriya. Nauka, Moscow

Newman MEJ, Barabási AL, Watts DJ (2006) The structure and dynamics of networks. Princeton Studies in Complexity Princeton University Press, Princeton, NJ

AQ4 Payette N (2011) Agent-based models of science. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 4, pp xx–xx

Price DJ de Solla (1963) Little science, big science. Columbia University Press, New York, NY

AQ5 Radicchi F, Fortunato S, Vespignani A (2011) Citation networks. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 7, pp xx–xx

Sandstrom PE (1994) An optimal foraging approach to information seeking and use. Libr Q 64(4):414–449 (stable JSTOR URL: http://www.jstor.org/stable/4308969)

Scharnhorst A (1988) Philosophische, methodologische und erkenntnistheoretische Probleme der Anwendung von Instrumentarien der Physik der Selbstorganisation auf quantitative Aspekte der Wissenschaftsentwicklung. Dr. phil. (PhD) Thesis. Humboldt-Universität zu Berlin, Berlin

Scharnhorst A, Garfield E (2010) Tracing scientific influence. Dyn Socio-Econ Syst Int J 2(1):1–33, available online at the URL: http://www.dyses.org.ar/ojs2333/index.php/jdyses/article/view/11, also available as arXiv preprint arXiv:1010.3525

Shiffrin RM, Börner K (2004) Mapping knowledge domains. Proc Natl Acad Sci USA 101(Suppl 1):5183–5185, (DOI: 10.1073/pnas.0307852100), also available online at the URL: http://ivl.slis.indiana.edu/km/pub/2004-shiffrin-mapknow.pdf. The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci U S A 101(Suppl 1):5183–5310

Small H, Griffith BC (1974) The structure of scientific literatures I: Identifying and graphing specialties. Sci Stud 4(1):17–40 (DOI: 10.1177/030631277400400102, stable JSTOR URL: http://www.jstor.org/stable/284536)

Snijders TAB, Steglich CEG, Schweinberger M, Huisman K (2008) Manual for SIENA version 3.2. University of Groningen, ICS, Department of Sociology, Groningen; University of Oxford, Department of Statistics, Oxford, available online at the URL: http://stat.gamma.rug.nl/s_man317.pdf and at the URL: http://www.stats.ox.ac.uk/ snijders /siena/s_man312.pdf

Stauffer D, Solomon S (2007) Ising, Schelling and self-organising segregation. Eur Phys J B 57(4):473–479 (DOI: 10.1140/epjb/e2007-00181-8), also available as arXiv preprint arXiv:physics/0701051

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. Technol Forecast Soc Change 28(2):93–122 (DOI: 10.1016/0040-1625(85)90009-5)

van den Besselaar P, Börner K, Scharnhorst A (2011) Science policy and the challenges for modeling science. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science AQ6 dynamics. Springer, Berlin, Ch 8, pp xx–xx

Vitanov NK, Ausloos MR (2011) Knowledge epidemics and population dynamics models for describing idea diffusion. In: Scharnhorst A, Börner K, van den Besselaar P (eds) Models of science dynamics. Springer, Berlin, Ch 3, pp xx–xx

Yablonskiĭ AI (1986) Matematicheskie modeli v issledovanii nauki. Nauka, Moscow

AUTHOR QUERIES

AQ1. Please update ref. Hanauske (2011)
AQ2. Please update ref. Lucio-Arias and Scharnhorst (2011)
AQ3. Please update ref. Mali et al. (2011)
AQ4. Please update ref. Payette (2011)
AQ5. Please update ref. Radicchi et al. (2011)
AQ6. Please update ref. van den Besselaar et al. (2011); Vitanov and Ausloos (2011)

Acknowledgements

The editors would like to acknowledge the events and contributions that helped shape this book project. In October 2009, they organized an international workshop on “Modelling Science: Understanding, Forecasting, and Communicating the Science System” in Amsterdam, The Netherlands, that brought together science modelers from different areas of science and inspired this book.[1] Two additional author meetings took place in Amsterdam in June 2010 and October 2010.[2] We are grateful to the authors for their willingness to communicate across disciplinary boundaries. We appreciate the time and effort of reviewers – chapter authors as well as external experts – who helped to enhance the quality and coverage of this book. Special thanks go to Kevin W. Boyack, Annamaria Carusi, Werner Ebeling, Serge Galam, Nigel Gilbert, Yves Gingras, Adam Lipowski, Staša Milojević, Frank Page, Sid Redner, and Henry Small (in alphabetical order). Matthias Hanauske took on the time-consuming task of final layouting the book in LaTeX. Thanks to his expertise, dedication and meticulousness all contributions turned into one book. A special thanks goes to Todd Nicholas Theriault for his very careful language editing, and to Daria for special support. We hope that this book will help readers gain a global understanding of models of science in existence so far and inspire future work on modeling the complex dynamics of the sciences.

This work was supported by the COST (European Cooperation in Science and Technology) action MP0801 – “Physics of Competition and Conflicts,” the Royal Netherlands Academy of Arts and Sciences (KNAW) *Strategiefonds* project Knowledge Space Lab, and the KNAW programme “Virtual Knowledge Studio for the Humanities and Social Sciences”, the James S. McDonnell Foundation, the Cyberinfrastructure for Network Science Center at Indiana University, USA, and the Rathenau Institute, The Netherlands.

---

[1] See http://modelling-science.simshelf.virtualknowledgestudio.nl/content/welcome for more material about the workshop.

[2] Materials from this workshop can be found here: http://mod_know.virtualknowledgestudio.nl/.

Contents

Contributors

Marcel Ausloos 7 rue des Chartreux, B-4122 Plainevaux, Belgium;
Previously at GRAPES, SUPRATECS, University of Liege, B-4000 Liege, Belgium, marcel.ausloost@ulg.ac.be

Marcel Ausloos is a Belgian engineer (ULG 1967). He turned to theoretical condensed matter for his M.Sc. (Brown U. 1970) and leaned toward statistical physics for his Ph.D. (Temple U. 1973). Thereafter, he has been applying statistical physics ideas and methods in various fields, ranging from biophysics, financial and econophysics, meteorology, linguistics, sociophysics, and many other non-equilibrium time-dependent systems. After some time in Berlin, at FUB, he rejoined the ULG, where he did whole his career, mainly teaching and performing some research. He founded the SUPRAS group of scientists belonging to various ULG laboratories with fundamental or applied science interests, with the goal of studying and understanding the properties of ceramics superconductors within a fundamental approach but motivated by potential applications. Later the center was expanded to become SUPRATECS (Services of the University Performing Research and Applications in the Technology of materials: Electroceramics, Composite and Superconductors). In parallel he initiated the GRASP, from which he was ejected, whence reforming the GRAPES (Group of Researchers for Applications of Physics in Economy and Sociology). He has contributed to the statistical mechanics theory through the study of spin systems, liquids, powders, and various disordered systems, – e.g., which can be described along fractal and multi-fractal ideas. He has been much interested in describing non-linear and out-of-equilibrium phenomena, – e.g. like growth and fracture, as well as transport phenomena, in particular under a temperature gradient or a magnetic field. He has a very long list of publications, edited books, organized many international conferences, presented many scientific contributions at meetings or gave seminars, and taught in many "schools". He has been taken as an expert for many projects by various financially granting bodies, has led many research programs, and been recognized through several honoring distinctions.

Peter van den Besselaar Department of Organization Science & Network Institute, VU University Amsterdam, The Netherlands, p.a.a.vanden.besselaar@vu.nl

Peter van den Besselaar's research interests are the dynamics of science, technology and innovation; science and innovation policy; users and design; and the effects of ICT on social change. He set up a group called "Science System Assessment" at the Rathenau Institute in the Netherlands. As a relative novelty for science and technology studies, the research program combined qualitative approaches with bibliometrics and modeling. Also, in his current position as professor for organization sciences at the Vrije Universiteit Amsterdam, his research focuses on the organization and dynamics of the science system, and on the development of e-social science methods and tools for science studies.

Katy Börner Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, U.S.A., katy@indiana.edu

Katy Börner's research focuses on the development of data analysis and visualization techniques for information access, understanding, and management. She is particularly interested in the study of the structure and evolution of scientific disciplines; the analysis and visualization of online activity; and the development of cyberinfrastructures for large-scale scientific collaboration and computation. She is a curator of the Places & Spaces: Mapping Science exhibit (http://scimaps.org) (Börner 2010).

Kevin W. Boyack SciTech Strategies, Albuquerque, NM, kboyack@mapofscience.com

Kevin W. Boyack's research is focused on developing detailed and accurate global maps of science. In addition, he has created algorithms to generate institution and country competency maps based on the SciTech global mapping techniques, and has done large scale comparisons of similarity approaches. Additional interests include the use of maps and metrics for research planning and evaluation.

Patrick Doreian Department of Sociology, University of Pittsburgh, 2602 WWPH, Pittsburgh, PA 15260, USA and Faculty of Social Sciences, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia, pitpat@pitt.edu

Patrick Doreian is Professor Emeritus of Sociology and Statistics at the University of Pittsburgh and a Research Faculty member of the Faculty od Social Sciences at the University of Ljubljana. His research focus is on the structure, operation and evolution of social networks. Substantively, his work includes studies of national scientific systems, voting patterns of the United Nations, decision making of the US Supreme Court, international migration patterns of professional football (soccer) players, network formation in police academies and environmental social movements. Methodologically, his interests include generalized blockmodeling of unsigned and signed relations, coupling blockmodels and exponential random graph models, and network autocorrelation models. Currently he is a co-editor of Social Networks and was a long time editor of the Journal of Mathematical Sociology.

Anuška Ferligoj Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia, Anuska.ferligoj@fdv.uni-lj.si

Anuska Ferligoj is professor of statistics at the University of Ljubljana. Her research focuses on multivariate analysis (especially cluster analysis without and with constraints, multicriteria clustering), on social network analysis (especially measuring networks and blockmodeling), on assessing reliability and validity of measurement and the use of statistical methods in various scientific disciplines. She is a co-editor of the journal "Advances in Methodology and Statistics" (Metodoloski zvezki) and member of several editorial boards of international scientific journals.

Santo Fortunato Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133 Torino, Italy, santo.fortunato@gmail.com

Santo Fortunato got his University degree at the University of Catania (Italy) in 1995 and his PhD in Germany, at Physics Department of the University of Bielefeld, in 2000. After four further years in Bielefeld as a postdoctoral research associate, he moved from statistical field theory to complex systems and joined Alessandro Vespignani's group at the School of Informatics of Indiana University where he stayed as a postdoc from 2005 till 2007. In 2007 he moved to the Institute for Scientific Interchange of Torino, where he became research leader in 2009. His research interests span from the theory and modeling of networks to social dynamics.

Matthias Hanauske Institute of Information Systems, Goethe-University, Grüneburgplatz 1, 60323 Frankfurt/Main. Germany, hanauske@wiwi.uni-frankfurt.de

Matthias Hanauske, obtained a doctoral degree in Physics (2004: Dr. phil. nat. (philosophiae naturalis)) and in Economics (2011: Dr. rer. pol. (rerum politicarum)) of the Goethe-University in Frankfurt/Main. His research topics ranges from general relativity, astro- and elementary particle physics to quantum game theory and evolutionary dynamics.

Luka Kronegger Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia, luka.kronegger@fdv.uni-lj.si

Luka Kronegger is a researcher and teaching assistant of Statistics at the Faculty od Social Sciences at the University of Ljubljana. His work includes studies of collaboration networks. He is oriented to methodological perspectives of modeling of network dynamics.

Diana Lucio-Arias Colombian Observatory of Science and Technology, carrera 15 no 37-59, Bogotá, D.C, Colombia, dlucio@ocyt.org.co

Diana Lucio-Arias, obtained a PhD in Social and Behavioral Sciences at the University of Amsterdam in 2010, currently she is the coordinator of research in innovation in the Colombian Observatory of Science and Technology (OCyT) where she is also involved in editing the book series Science and Technology Indicators in Colombia. Diana has experience in topics related to scientometric indicators, complexity in social systems and innovation systems.

Franc Mali Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia, franc.mali@fdv.uni-lj.si

Franc Mali's research focuses on social studies of science and technology. He mainly works on the social and ethical aspects of modern scientific development. He is particularly oriented to the issues of the functioning of the modern communities of scientists, intermediary scientific structures, evaluation of scientific productivity and social (policy) regulation of the new emerging sciences.

Staša Milojević SLIS, Indiana University, Bloomington, IN, smilojev@indiana.edu

Staša Milojević's research focuses on studying how modern scientific disciplines/fields form, organize and develop. She approaches modern scientific fields/disciplines as complex heterogeneous socio-cultural networks of people, ideas, documents and institutions. In large-scale longitudinal studies of fields such as nanotechnology she combines models, theories and methods from information science, science and technology studies, and social network analysis.

Steven Morris Baker Hughes, Houston, Texas, Steven.Morris@bakerhughes.com

Steven A. Morris is an electrical engineer and native of Tulsa, Oklahoma. He earned B.S.E.E. and M.S.E.E. degrees from Tulsa University in 1983 and 1986 respectively, and worked as a research scientist in the petroleum industry for 15 years. Returning to academic studies in 1999, he conducted research in the mapping of journal and patent literature for technology forecasting and technology management applications. After earning a Ph.D. degree in Electrical Engineering from Oklahoma State University in 2005, he developed crossmaps as a means to visually examine and communicate correlations among the diverse scholarly entities and their attributes.

Nicolas Payette Philosophy Department, Université du Québec à Montréal, 455 René-Lévesque Est, Montréal, Québec, Canada, H2L 4Y2, nicolaspayette@gmail.com

Nicolas Payette is working on problems in the philosophy of science using techniques from the fields of agent-based modeling, evolutionary computation, artificial intelligence and social network analysis. His main premise is that science is some sort of distributed cognitive system that can be reverse engineered (and hopefully optimized) using these methods. His current project aims at simulating the evolution of scientific theories by having agents connected in collaboration networks explore an epistemic landscape by producing, exchanging and modify ideas represented by points in this landscape.

Filippo Radicchi Howard Hughes Medical Institute (HHMI), Northwestern University, 2145 Sheridan Road, Evanston, IL, 60208 USA, f.radicchi@gmail.com

Filippo Radicchi is a Research Specialist of Howard Hughes Medical Institute (HHMI) at Northwestern University. He got his M.Sc. degree in Physics in 2003 at the University of Rome "Tor Vergata" (Italy), and he received his Ph.D. in Physics in 2007 at Jacobs University Bremen (Germany). He worked between 2007 and 2010 as Research Scientist at Institute for Scientific Interchange (ISI) Foundation in Torino (Italy), and he worked in 2010 as Postdoctoral Fellow at the department of Chemical and Biological Engineering at Northwestern University (USA). Filippo Radicchi's research focuses on the study of complex systems, social networks and citation analysis.

Andrea Scharnhorst Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences, Anna van Saksenlaan 10, 2593 HT Den Haag, The Netherlands, andrea.scharnhorst@dans.knaw.nl

Andrea Scharnhorst's research focuses on modelling and simulating the emergence of innovations. Hereby, innovation is understood in a broad sense including new modes of behavior and learning, forms of communication, technologies or scientific ideas. She has work at the interface between physics, social sciences and humanities. She is particularly interested to better understand how to operationalize the concept of a knowledge landscape, how to translate it into different mathematical models, how to visualize the landscape and the individual trajectories in it, and how to better understand the framing conditions and the internal dynamic mechanism of the self-organizing science system.

Alessandro Vespignani Center for Complex Networks and Systems Research, School of Informatics and Computing Indiana University, 919 East 10th St., Room Bloomington, IN, 47408 USA, alexv@indiana.edu

Alessandro Vespignani is currently James H. Rudy Professor of Informatics and Computing at Indiana University where he is the director of the Center for Complex Networks and Systems Research (CNetS) and associate director of the Pervasive Technology Institute. Recently Vespignani's research activity focuses on the interdisciplinary application of statistical and simulation methods in the analysis of epidemic spreading phenomena and the study of biological, social and technological networks. Vespignani is scientific director of the Institute for Scientific Interchange Foundation, elected fellow of the American Physical Society and is serving in the board/leadership of several scientific organizations and journals.

Nikolay K. Vitanov Institute of Mechanics, Bulgarian Academy of Sciences, Akad. G. Bonchev Str., Bl. 4, 1113 Sofia, Bulgaria, vitanov@imbm.bas.bg

Nikolay K. Vitanov is full professor, vice-managing director and director for science and corporate relations of the Institute of Mechanics of the Bulgarian Academy of Sciences. His scientific research is concentrated around the area of complex systems behaviour. In more detail he works in the area of mathematical social dynamics, econophysics, nonlinear time series analysis, theory of dynamical systems, theory of turbulence and theory of nonlinear partial differential equations. Author of more than 100 scientific papers and co-author of two books in Bulgarian: Population dynamics and national security (2005), and Social dynamics without formulas (2008).

Part I
Foundations

Chapter 1
An Introduction to Modeling Science: Basic Model Types, Key Definitions, and a General Framework for the Comparison of Process Models

Katy Börner, Kevin W. Boyack, Staša Milojević, and Steven Morris

1.1 Introduction

Science is in a constant state of flux. Indeed, one of the purposes of science is to continually generate new knowledge, to search for or create the next breakthrough that will open new doors of understanding. Science can also be viewed as a research process in which scholars coordinate their actions, working in a wide range of institutions and using ever better methods and instruments, to generate new knowledge, which is then recorded in tangible forms as journal articles, reports, books, patents, data, and software repositories, etc. (Whitley 1984).

Science is a complex phenomenon, and as such it captures the interest of a wide range of researchers in fields such as history, philosophy and sociology of science, and scientometrics. From the standpoint and for the purposes of scientometrics and modeling of science, science can be defined as a social *network of researchers* that generate and validate a *network of knowledge*. This definition is based on the premise that science consists of knowledge and ideas that are produced and validated by a community of researchers. Researchers belong to institutions that support activities related to scientific research and inquiry. The way knowledge is produced, organized, and disseminated is dependent on historical, institutional, political, and research contexts. At the same time, the meanings of the concepts

K. Börner (✉) · S. Milojević
AQ1 SLIS, Indiana University, Bloomington, IN
e-mail: katy@indiana.edu; smilojev@indiana.edu

K.W. Boyack
SciTech Strategies, Albuquerque, NM
e-mail: kboyack@mapofscience.com

S. Morris
Baker Hughes, Houston, Texas
e-mail: Steven.Morris@bakerhughes.com

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_1,

one uses to describe science and knowledge are not only constantly changing but are also culturally and historically specific. For example, in recent years, there is a tendency towards heterogeneous (interdisciplinary) teams of researchers solving pressing social problems with higher accountability (Gibbons et al. 1994; Nowotny et al. 2001). Due to the changing nature of knowledge and the changing social structure of science, some of the institutional forms and established practices in science are undergoing changes themselves.

The idea of studying science using scientific methods is at the core of scientometrics. Many scientometric studies describe the structure and evolution of science, while a few others aim to replicate and predict the structure and dynamics of science. It is the latter group that is the focus of this chapter and this book.

1.1.1 Science as a Social Activity

The relationships between scholars and the institutions they are affiliated with constitute the social characteristics of science. Scientific knowledge does not exist in a vacuum. It requires social infrastructure for support. This social infrastructure can be manifest in forms such as funding, oversight, management, collaboration, and less formal modes of communication. Researchers often work collaboratively to produce new knowledge. They also use both formal and informal channels to communicate their results. At the same time, they are embedded in a number of organizations and institutions, such as university departments, research centers, and research institutes. These institutions, together with meta-institutions such as government agencies, industry segments, or universities, shape rewards in science.

Different interactions in which scholars engage can result in different aggregates, such as invisible colleges, specialties, disciplines, and interdisciplines.[1] Studies of science as a social activity mostly focus on the stages of development of smaller units of aggregation, such as specialties. Studies that focus on the social aspects of science view science as a development of social structures, viewed qualitatively as stages of social group formation (Crane 1972; Wagner 2008), or quantitatively as stages of cluster formation (Palla et al. 2007).

The intricacies of the relationships between social and cognitive aspects of science are most visible among relatively small groups of scholars over short periods of times. At the same time, these scholars are embedded, through both training and employment, in larger units, such as fields or disciplines or university departments, which exercise significant power over rewards and thus shape the behavior of scholars.

---

[1] The terms "multidisciplinary", "interdisciplinary" and "transdisciplinary" have been used to describe research activities, problems, institutions, teachings, or bodies of knowledge, each with an input from at least two scientific disciplines.

1.1.2 Science as a Knowledge Network

The cognitive structure of science consists of ideas and relationships between ideas. Cognitive studies focus on science as a body of knowledge. There is no unanimously accepted definition of cognitive structure, and studies on the topic range from those dealing with epistemology, the structure of scientific theories, and the relationship between theoretical and empirical work, to the studies of the cognitive consensus among scientists. Given the importance of textual documents in the practice of science (Callon et al. 1983; Latour and Woolgar 1986), it is natural to focus on the shared conceptual system of scientific communities as expressed through the terminology used in those documents. In this paper, we focus on the studies of scientific knowledge using documents or artifacts produced by scholars as the data.

There are different ways in which one can study scientific knowledge using documents as a starting point. One approach is to study textual elements associated with the documents (e.g., words from titles, abstracts, keywords or index terms, or even full text) using, for example, word co-occurrence analysis. Another approach is to treat references as concept symbols (Small 1978) and then perform a whole range of analyses using references as a data source. These analyses can be used to produce maps of science that seek to visually describe the structure of the data (Börner et al. 2003). A third approach is to take journals as units of analysis and study their subjects. These analyses are often used for studying interdisciplinarity. Regardless of the approach, these studies focus mostly on the evolving structure of scientific ideas or the emergence, growth, and diffusion of scientific ideas. They are highly relevant for funding agencies that continually seek to support the most promising and/or emerging topics in science.

1.2 Science Models

This section introduces a general definition of science models and explains how they are designed. It then discusses different model types. This book focuses on quantitative predictive models that might be universal or concrete. Frequently, there is the desire to model a system at multiple levels.

1.2.1 Definition and General Design of a Science Model

"Model" is a word with a number of meanings. The *Oxford English Dictionary*, for example, states in one of its 17 definitions of the word that a model is "a simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.; a conceptual or mental representation of something."[2] In philosophy of science, models can be representations of certain phenomena or data, or they can represent a theory.[3] In the social sciences, models are simplified representations of an aspect of the real world. They are "a generic term for any systematic set of conjectures about real world observations."[4] In system science and applied mathematics, a model is "an encapsulation of some slice of the real world within the confines of the relationships constituting a formal mathematical system."[5]

Here, we are interested in models that capture the structure and dynamics of scientific endeavor to gain insights into the inner workings of science. *Structure* can be defined as a regular pattern in the behavior of elementary parts of a system based on observations of repeated processes of interaction. Typical time frames used in structural models can be as short as a month or as long as a decade. *Dynamics* refers to the processes and behaviors that lead to changes (e.g., birth, merge, split, or death) (Palla et al. 2007) in the structural units of science (e.g., research teams, specialties) or their interlinkages. Different model types are discussed in the next section. Recent work aims to develop models that describe the *interplay of structure and dynamics* to increase our understanding of how usage (e.g., collaboration of citation activity) impacts the structure of science and how structure supports activity.

In general, the study of science aims to answer specific questions such as when (temporal), where (spatial), what (topical), or with whom (network analysis), or combinations thereof. Temporal questions are commonly answered by dynamic models, including those based on linear regression and those that use sudden bursts of activity as an indicator of new developments. Spatial and topical questions assume an underlying geographic or semantic space and are often answered using structural models. They might simulate people's foraging for information, collaborators, or reputation in a model space analogous to that used by anthropologists to study food foraging. Other models adopt approaches from epidemiology to help us understand the impact of the origin of diffusing entities (tangible ones like people or intangible ones like ideas), infection/adoption rate, seasonality effects (e.g., papers published during spring semester or summer break), etc., on diffusion patterns and dynamics. In addition, there are models that simulate the growth of homogeneous or heterogenous networks, diffusion dynamics over networks, or the interplay of network structure and usage. Recent work in epidemiology aims to understand the interaction of epidemic spreading and social behavior (e.g., staying home when you

---

[2] *Oxford English Dictionary Online*, s.v. "model," accessed January 20, 2011, http://www.oed.com/view/Entry/120577?rskey=r3QCjg&result=1&isAdvanced=false.

[3] Roman Frigg and Stephan Hartmann, "Models in Science," *The Stanford Encyclopedia of Philosophy* (Summer 2009 Edition), ed. Edward N. Zalta, http://plato.stanford.edu/archives/sum2009/entries/models-science/.

[4] Charles A. Lave and James G. March, *An Introduction to Models in the Social Sciences* (Lanham: University Press of America, 1993), 4.

[5] John L. Casti, *Alternate Realities: Mathematical Models of Nature and Man* (New York: Wiley, 1989), 1.

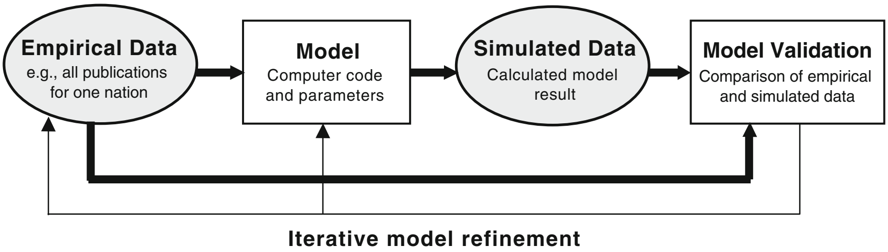


Fig. 1.1 General model design, validation, and refinement process

are sick). Analogously, it is desirable to study and model the effect of breakthrough ideas on scholarly network formation and usage.

Model design typically involves the formulation of a scientific hypothesis about the identification of a specific structure or dynamics. Often, this hypothesis is based on analysis of patterns found in empirical data. Whether the hypothesis is based on data or in theory, an empirical dataset needs to be available to test model results. Next, an algorithmic process is designed and implemented using either tools (e.g., NetLogo, RePast) or custom codes that attempt to mathematically describe the structure or dynamics of interest. Subsequently, simulated data are calculated by running the algorithm and validated by comparison with empirical data. Resulting insights frequently inspire new scientific hypotheses, and the model is iteratively refined or new models are developed. The general process is depicted in Fig. 1.1.

1.2.2 Qualitative Models vs. Quantitative Models

There are two major types of models: *Qualitative models* often use verbal descriptions of general behavior. *Quantitative models* express units of analyses, their interrelations and dynamics using properties susceptible of measurement. The latter are the focus of this book.

1.2.3 Deductive (Top-Down or Analytical) Models vs. Inductive (Bottom-Up or Synthetic) Models

Deductive models take a “top-down” approach by working from the more general to the more specific. For example, a deductive modeling approach might start with a general theory and then narrow it down into more specific hypotheses that can be tested. Deduction can be seen as the identification of an unknown particular based on the resemblance of the particular to a set of known facts.

Inductive models take a "bottom-up" approach that starts with specific observations and measurements, continues with the identification of patterns and regularities, formulates some tentative hypotheses that can be explored, and results in general conclusions or theories. Induction is also known as the formation of a generalization derived from examining a set of particulars. It is more open-ended and exploratory, especially at the beginning.

1.2.4 Deterministic Models vs. Stochastic Models

Deterministic models describe the behavior of an object or phenomenon whose actions are entirely determined by its initial state and inputs. In deterministic models, a given input will always result in the same output. A single estimate is used to represent the value of each model variable. Examples are physical laws (e.g., Newton's laws) that can be used to describe and predict planetary motion.

Stochastic (also called probabilistic) models make it possible to predict the behavior of an object or phenomenon if the influence of several unknown factors is sizable – the subsequent state is determined both by predictable actions and by a random element. They cannot predict the exact behavior but predict the probability that a particular value will be observed at a particular time within a known confidence interval. Ranges of values (in the form of a probability distribution) are used to describe each model variable.

1.2.5 Descriptive Models vs. Process Models

Quantitative models of science can be further divided into two categories: descriptive models and process models. Both can be used to make predictions. *Descriptive models* aim to describe the major features of typically static data sets. Results are communicated via tables, charts, or maps. The focus of this book is on *process models*, which aim to capture the mechanisms and temporal dynamics by which real-world networks are created (Newman and Leicht 2007; Zhang et al. 2010), with particular emphasis on identification of elementary mechanisms that lead to the emergence of specific network structures and dynamics. These models aim to simulate, statistically describe, or formally reproduce statistical characteristics of interest, typically by means of formulas or implemented algorithms. Formal mathematical approaches to process modeling work best for static, homogeneous worlds. Computational models, however, allow us to investigate richer, more dynamic environments with greater fidelity helping us to understand and explain the dynamic nature of science.

Note the difference between laws and computational models. Bibliometric *laws* are, in reality, descriptive models of data that are held true for certain classes of systems. Examples include Lotka's law (Lotka 1926), Bradford's law

(Bradford 1934), and Zipf's law (Zipf 1949). *Computational models* describe the structure of dynamics of science using different computational approaches such as agent-based modeling, population models (Bettencourt et al. 2008), cellular automata, or statistical mechanics.

A number of studies that use co-authorship networks to study network dynamics (Barabási et al. 2002; Barabási and Albert 1999; Farkas et al. 2002; Nagurney 1999; Newman 2001) reveal the existence of small-world and scale-free network topologies (see Sect. 1.2.6) and preferential attachment (Price 1976) as a structuring factor. Preferential attachment in the context of networks means that the well-connected nodes are more likely to attract new links.

1.2.6 Universal Models vs. Domain-Specific Models

Models can be designed at different levels of generality or universality. *Universal models* aim to simulate processes that hold true across different domains and datasets. Examples include scale-free network models (Barabási and Albert 1999) or small-world network models (Watts and Strogatz 1998) generating network structures that can be found in vastly diverse systems such as social, transportation, or biological networks. *Domain-specific models* aim to replicate a concrete dataset in a given domain. One example is Goffman's (Goffman 1966) application of an epidemic model to study the diffusion of ideas and the growth of scientific specialties. By using mast cell research as a case study, he demonstrated that it was possible to see growth and development as sequences of overlapping epidemics. In this and in other dynamic models, one simulates the dynamic properties of the system by applying certain global laws characteristic of complex systems. This is particularly useful for modeling the growth of a whole system, some part of a system, or of a measure that corresponds to a size. Price studied the growth of science using data until about 1960 and observed an exponential growth (Price 1963). Since then, growth has been largely linear, mirroring the massive but linear growth in R&D funding.

Today, it is assumed that there are two ways science can grow: homogeneously and heterogeneously. Homogeneous growth is a simple expansion of a given unit. Heterogeneous growth, on the other hand, means differentiation or rearrangement of component elements. Highly differentiated, heterogeneous growth of science can be viewed through authorship patterns. For example, not only is the number of authors per paper increasing over time, but also these authors come from different disciplines, different institutions, and different knowledge-production sites (e.g., universities and industries). In addition, there is a wide geographic distribution of co-authors as well. This is the result of the globalization of science and the role that specialized knowledge plays in the development of science. A particularly promising area of research is the study of co-evolving networks of co-authors and paper-citations (Börner et al. 2004), as well as work that examines the interplay of

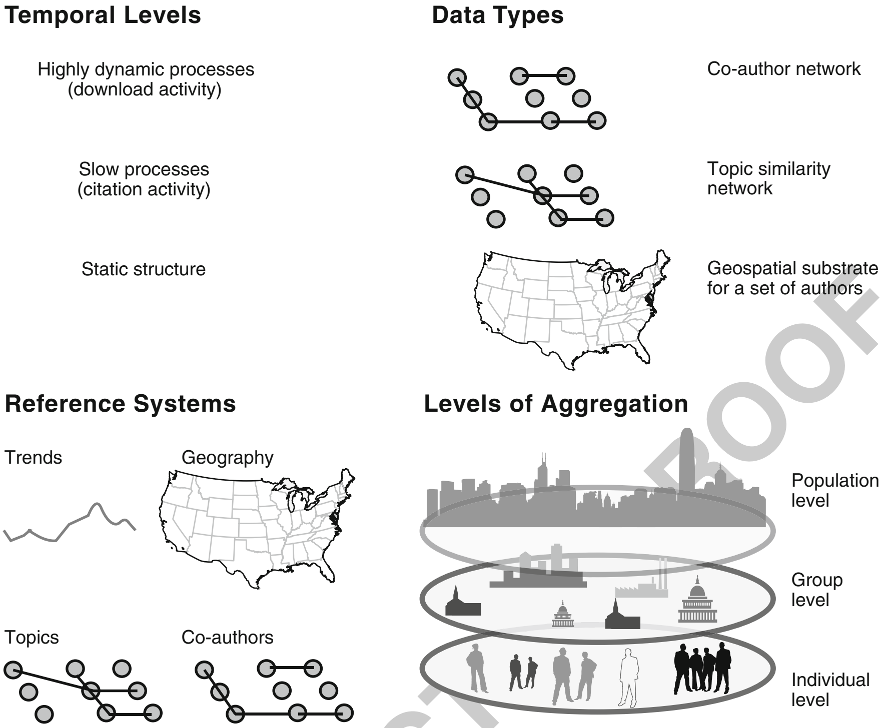


Fig. 1.2 Temporal levels (*top left*), data types (*top right*), reference systems (*lower left*), and levels of aggregation (*lower right*)

existing network structures and resulting scholarly dynamics that, in turn, affect the growth of scholarly networks.

1.2.7 Multi-Level and Multi-Perspective Models

It is often desirable to model a system at multiple levels using different vantage points (see Fig. 1.2).

For example, the different levels could represent:

- Temporal scales – different levels describe the structure and/or dynamics of a system at different points in time.
- Data types – different levels represent different relations/dynamics for the very same set of elements (e.g., co-author, co-PI, co-investigator, co-inventor, author co-citation, and topical similarity for a set of nodes).

- Reference systems – different levels provide different views of the same data (e.g., a map of NIH funding is linked to a map of authors is linked to a map of their MEDLINE publications).
- Levels of aggregation – levels might represent different geospatial aggregations, topical aggregations, or network aggregations such as individual, group, population level data, e.g., co-author networks, research communities, or invisible colleges.

1.2.8 Exemplification Using Predictive Workflows

As mentioned in Sect. 1.2.1, models of science aim to answer when, where, what and with whom questions at different levels of aggregation, e.g.,

- *when (temporal):* days, weeks, months, years, decades, centuries; several journal volumes/issues make up years
- *where (spatial):* postal codes, counties, states/provinces, countries, continents. NanoBank has an elaborate system for this. Congressional districts matter. OPEC countries, EU, etc., aggregations of countries
- *what (topical):* terms make up topics, documents, and lines of research; papers appear in journals, journals group into disciplines or subject categories, major fields, or all of science
- *with whom (network analysis):* person is part of a research team, part of a research community/invisible college; person works at an institution, institution is part of a sector (e.g., academia, government, industry).

Answers to these different types of questions each demand their own data structures (e.g., time-stamped data or networks). Below, we provide sample modeling workflows that aim to answer research or science policy questions.

Although models of science aim to answer the when, where, what, and with whom questions mentioned above, it is important to relate them to the needs of science policy and practice. There are many types of questions currently being asked by decision-makers (from team leaders to university officials to agency heads) that can potentially be informed by science models. These include:

- How do changing resources alter the structure of science (at multiple levels of aggregation)? What areas would benefit most from increased funding?
- What science is currently emerging or likely to emerge in the near future?
- How can I create or strengthen a particular R&D area at my institution? What key expertise and resources are needed?

To a large degree, science policy and practice is interested in models as a way to make informed decisions regarding future (investment) strategies in science. In that respect, they are interested in predictive models of science.

To date, the majority of predictive models have sought to describe phenomena at high levels of aggregation. Descriptive models have much more often been able to describe phenomena at very detailed levels. What is needed in the future is a merging of the scales that are currently possible using descriptive models with the predictive power of computational models. This combination has an unparalleled opportunity to impact science policy and the practice of science in very significant ways. We would like to extend this as a challenge to the science-modeling community. To illustrate this challenge and this opportunity, we provide an example of how such a model (or combination of models) could be used to provide answers to detailed questions.

Dynamics of the S&T system. It is well known that topics in science are born, can merge or split, and eventually die. Some descriptive models can show the past dynamics of topics or disciplines; isolated studies have examined this issue in some segments of the literature. Predictive models have reproduced the growth characteristic of the life span of many scientific fields (Gupta et al. 1997). However, to date, there has been no comprehensive study to (1) track communities or specialties over all of science to discover the empirical birth, merge, split, and death rates (the comprehensive descriptive model), and (2) to correlate those rates with properties of the communities or specialties (the comprehensive predictive model). This combination could result in a highly specific model that could be used to predict (based on model parameters fit to past performance) the status of each current community for the next several years. Such a predictive model would be an extremely powerful tool for decision-makers.

1.3 Basic Conceptualization and Science-Modeling Terminology

Despite the fact that different science models have been designed to answer vastly different questions at many levels of generality, the discussion above has implicitly assumed, without explicitly stating, that any model of science must be based on some sort of framework or conceptualization of science, its units, relationships, and processes. In an attempt to provide a unifying conceptualization (Börner and Scharnhorst 2009) for the comparison of models, we present here two different frameworks, one starting with terms and definitions, and one starting with a visual network approach. The two frameworks have a high degree of overlap, and demonstrate that useful frameworks can be approached from multiple perspectives. There are some facets of these frameworks that are similar to those previously published by Morris and Rodriguez (Morris and van der Veer Martens 2008; Morris and Yen 2004; Rodriguez et al. 2007). However, there are many differences as well.

The origin, usage, and utility of key terms very much depends on the goal and type of modeling performed. Models that conceptualize science as a social activity (see Sect. 1.1.1) will use researchers, teams, and invisible colleges as key *social terms*. Models that simulate science as a knowledge network (see Sect. 1.1.2) have to define *knowledge terms* such as documents and journals. Models that place a central role on the bibliographic data used in model validation require a definition of *bibliographic terms*. Models that conceptualize science as an evolving system of co-author, paper-citation, and other networks will need to define *network terms*. Other models aim to capture the phenomenology of science or try to provide actionable knowledge for science policy decisions and hence define *phenomenological terms* and *policy/infrastructure terms*. The majority of the underlined terms are defined in the Appendix; the definitions provide more information on the concrete interlinkages between terms. Exemplary sets of essential terms (concepts) are given here:

- *Social terms:* researcher, team, invisible college, research community, specialty, institution, collaboration.
- *Knowledge terms:* base knowledge, line of research, discipline, field of study, research front, communication, knowledge diffusion, knowledge validation.
- *Bibliographic terms:* author, document (e.g., article, patent, grant), reference, citation, journal, term, topic.
- *Network terms:* network, node, link, clustering, network metric.
- *Phenomenological terms:* core and scatter, hubs and authorities, aggregation, overlap, distributions, bursts, drifts, trends.
- *Policy/Infrastructure terms:* funding, indicator, metrics.

Note that there are strong interrelations among these terms within and across the different term sets:

- Most researchers are authors.
- References and citations are links between papers.
- Researchers aggregate to teams, invisible colleges, research communities; they are affiliated with an institution.
- Journals include papers; papers have references and might be cited; papers are comprised of terms and address a specific topic.
- Clustering occurs not only in networks but also over time (e.g., only authors who are alive can co-author) and geospatial and topic space (e.g., authors who are geospatially close and work on similar topics are more likely to co-author).

The most inconsistently used terms are those used to describe

- *Social groupings* such as invisible colleges, research community, specialty and
- *Knowledge groupings* such as line of research, field of study, discipline.

Authors of the book chapters were encouraged to conform to or redefine the definitions given in the Appendix. Readers of the book might like to do the same.

Note that many different groupings of these terms are possible. Leydesdorff (Leydesdorff 1995) suggested a three-dimensional space of different units of analysis: social dimensions (people, institutions), institutional dimensions (rules, funding, metrics, indicators), and cognitive dimensions (texts, journals), see Fig. 1.3. The three derivative two-dimensional spaces represent different lines of research.

- Social x institutional dimensions: Sociology of science
- Social x cognitive dimensions: Scientometrics, informetrics
- Institutional x cognitive dimensions: Philosophy of science, artificial intelligence

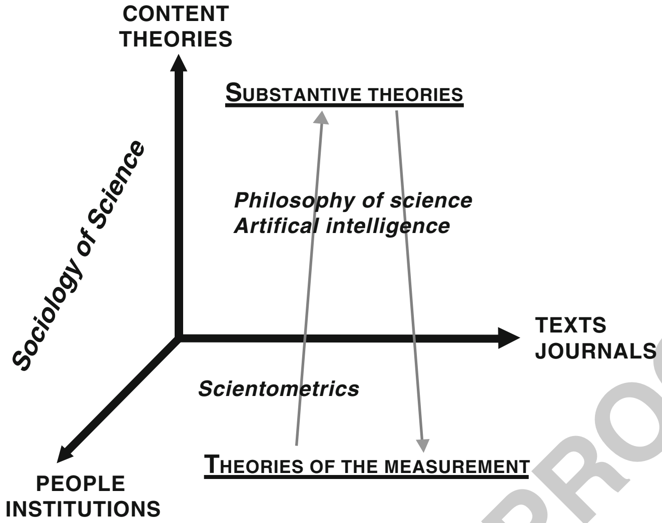


Fig. 1.3 Three-dimensional space by Leydesdorff (Leydesdorff 1995). The three axes stand for units of analysis. A phenomenon can be represented as a point in this space with a value on each of the three axes via projection. For example, an institutional rule that would be attributed to an institution might be represented as text, and has cognitive (substantive) content

In an analogy to a physical system, social dimensions are the "volume," cognitive dimensions are the "temperature," and institutional dimensions are the "pressure."

A system-theoretic approach by sociologist Luhmann (1995) depicts science as a self-organizing process within society that takes human resources, education, and funding as input and produces papers, books, patents, and innovations as output. While science strives for "truth," economy aims for profit.

A final alternative, network-based approach is given in Fig. 1.4. This conceptualization is useful when developing models for science policy-makers with a deep interest in indicators. Here, social, knowledge, and topical descriptor networks are extracted to study base entities and their physical aggregations into teams, institutions, journals, and documents. Conceptual aggregations such as invisible colleges, specialties, or smaller communities can be analyzed and mapped, and can show signs of incremental growth, emergence, and breakthrough, or controversy and conflict, depending on the actual dynamics of the science involved. Temporal changes in lines of research or bursts and drifts in time-stamped texts can be calculated and modeled. The ultimate goal is the support of effective funding, communication, collaboration, and their validation.

We note that many different conceptualizations of science are possible, and that the two presented here are only examples. They are not intended to provide an exhaustive list of the units of science that can be analyzed, but rather to suggest that one should be able to place the units and interactions used in any model of science in a coherent framework that will be useful to others.

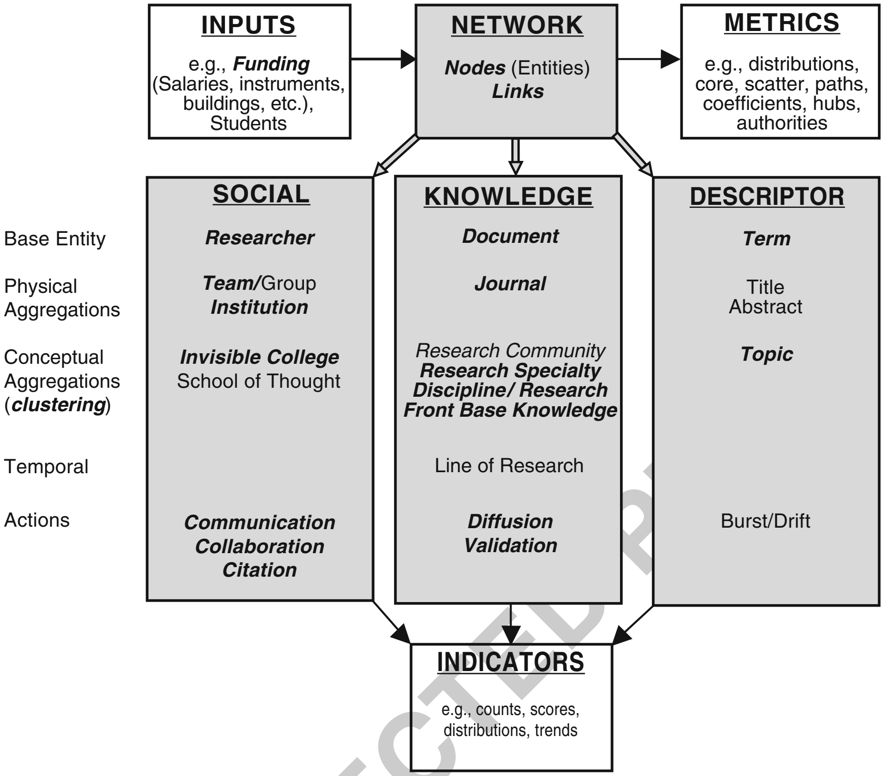


Fig. 1.4 Network-centric grouping of key terms used in models relevant for science policy making. Terms in bold italic are defined in the Glossary

1.4 Overview of Major Science Models

The remainder of this book reviews major process models that were developed in many different areas of science. Among them are

- *Statistical approaches and models* which are "based on the laws and distributions of Lotka, Bradford, Yule, Zipf-Mandelbrot, and others [and] provide much useful information for the analysis of the evolution of systems in which development is closely connected to the process of diffusion of ideas" (Chap. 3, p. 1);
- *Deterministic dynamical models* that are "considered to be appropriate for the analysis of [evolving] 'large' societal, scientific and technological systems for the case when the influence of fluctuations is not significant" (Chap. 3, p. 1);
- *Stochastic models* which are "appropriate when the system of interest is 'small' but when the fluctuations become significant for its evolution" (Chap. 3, p. 1);
- *Agent-based models (ABM),* which "are concerned with the micro-level processes that give rise to observable, higher-level patterns. If an ABM can generate

some macrophenomenon of interest, then it can at least be considered a candidate explanation for it." (Chap. 4, p. 6)

- *Evolutionary game theory (EGT)* is "a time-dependent dynamical extension of 'Game Theory' (GT), which itself attempts to mathematically capture behavior in strategic situations in which an individual's success in making choices depends on the choices of others. EGT focuses on the strategy evolution in populations to explain interdependent decision processes happening in biological or socio-economic systems" (Chap. 5, p. 2);
- *Quantum game theory* is "a mathematical and conceptual amplification of classical game theory (GT). The space of all conceivable decision paths is extended from the classical measurable strategy space in the Hilbert space of complex numbers. Through the concept of quantum entanglement, it is possible to include a cooperative decision path caused by cultural or moral standards" (Chap. 5, p. 18).

Figure 2.1 in Chap. 2 sketches the temporal evolution of the different model types. Chapters 3–7 each feature a table that lists major models reviewed in that chapter. While Chaps. 3–5 each review one specific model type, Chaps. 6 and 7 discuss different types of models that address questions related to the structure and dynamics of co-author and paper-citation networks respectively.

Acknowledgements We would like to thank the chapter authors for their constructive comments and expert input. This work is funded in part by the James S. McDonnell Foundation, the National Science Foundation under award SBE-0738111, and the National Institutes of Health under award NIH U24RR029822.

References

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439): 509–512 (DOI: 10.1126/science.286.5439.509), also available as arXiv preprint arXiv:cond-mat/9910332

Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311(3–4):590–614 (DOI: 10.1016/S0378-4371(02)00736-7), also available as arXiv preprint arXiv:cond-mat/0104162

Bettencourt LMA, Kaiser DI, Kaur J, Castillo-Chávez C, Wojick DE (2008) Population modeling of the emergence and development of scientific fields. Scientometrics 75(3):495–518 (DOI: 10.1007/s11192-007-1888-4)

Börner K, Chen C, Boyack KW (2003) Visualizing knowledge domains. Ann Rev Inf Sci Technol 37:179–255 (DOI: 10.1002/aris.1440370106)

Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. Proc Natl Acad Sci USA 101(Suppl 1):5266–5273 (DOI: 10.1073/pnas.0307625100), also available as arXiv preprint arXiv:cond-mat/0311459. The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci USA 101(Suppl 1):5183–5310

Börner K, Scharnhorst A (2009) Visual conceptualizations and models of science. Editorial for the special issue 'Science of science: Conceptualizations and models of science'.

J Informetr 3(3):161–172 (DOI: 10.1016/j.joi.2009.03.008), also available as arXiv preprint arXiv:0903.3562

Bradford SC (1934) Sources of information on specific subjects. Eng Illustrated Weekly J 137(3550):85–86. Reprinted: (1985) J Inform Sci 10(4):176–180 (DOI: 10.1177/016555158501000406)

Callon M, Courtial JP, Turner WA, Bauin S (1983) From translations to problematic networks: An introduction to co-word analysis. Soc Sci Inf 22(2):191–235 (DOI: 10.1177/053901883022002003)

Crane D (1972) Invisible colleges: Diffusion of knowledge in scientific communities. University of Chicago Press, Chicago, IL

Farkas I, Derényi I, Jeong H, Neda Z, Oltvai ZN, Ravasz E, Schubert A, Barabási AL, Vicsek T (2002) Networks in life: Scaling properties and eigenvalue spectra. Physica A 314(1–4):25–34 (DOI: 10.1016/S0378-4371(02)01181-0), also available as arXiv preprint arXiv:cond-mat/0303106

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M (1994) The new production of knowledge: The dynamics of science and research in contemporary societies. Sage Publications, London

Goffman W (1966) Mathematical approach to the spread of scientific ideas – the history of mast cell research. Nature 212(5061):449–452 (DOI: 10.1038/212449a0)

Gupta BM, Sharma P, Karisiddappa CR (1997) Growth of research literature in scientific specialities. A modelling perspective. *Scientometrics* 40(3):507–528 (DOI: 10.1007/BF02459297)

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago, IL

Latour B, Woolgar S (1986) Laboratory life: The construction of scientific facts. Princeton University Press, Princeton, NJ

Leydesdorff L (1995) The challenge of scientometrics: The development, measurement, and self-organization of scientific communications. Science Studies/Wetenschapsstudies. DSWO Press, Leiden; Leiden University, Leiden

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–323

Lucio-Arias D, Leydesdorff L (2009) The dynamics of exchanges and references among scientific texts, and the autopoiesis of discursive knowledge. J Informetr 3(3):261–271 (DOI: 10.1016/j.joi.2009.03.003), also available as arXiv preprint arXiv:0911.1308

Luhmann N (1995) Social systems. Stanford University Press, Stanford, CA. Translation of: Luhmann N (1984) Soziale Systeme: Grundriß einer allgemeinen Theorie. Suhrkamp, Frankfurt am Main

Merton RK (1973) The sociology of science: Theoretical and empirical investigations. University of Chicago Press, Chicago, IL

Morris SA, van der Veer Martens B (2008) Mapping research specialties. Ann Rev Inf Sci Technol 42(1):213–295 (DOI: 1002/aris.2008.1440420113)

Morris SA, Yen GG (2004) Crossmaps: Visualization of overlapping relationships in collections of journal papers. Proc Natl Acad Sci USA 101(Suppl 1):5291–5296 (DOI: 10.1073/pnas.0307604100). The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci USA 101(Suppl 1):5183–5310

Nagurney A, Siokos S (1999) Financial networks: Statics and dynamics. Advances in Spatial Science. Springer, Berlin

Newman MEJ (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64(2):025102(R), 4 pp (DOI: 10.1103/PhysRevE.64.025102), also available as arXiv preprint arXiv:cond-mat/0104209

Newman MEJ, Leicht EA (2007) Mixture models and exploratory analysis in networks. Proc Natl Acad Sci USA 104(23):9564–9569 (DOI: 10.1073/pnas.0610537104), also available as arXiv preprint arXiv:physics/0611158

Nowotny H, Scott P, Gibbons M (2001) Re-thinking science: Knowledge and the public in an age of uncertainty. Polity Press, Cambridge

Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446(7136):664–667 (DOI: 10.1038/nature05670), also available as arXiv preprint arXiv:0704.0744

Price DJ de Solla (1963) Little science, big science. Columbia University Press, New York, NY

Price DJ de Solla (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306 (DOI: 10.1002/asi.4630270505), also available online at the URL: http://www.asis.org/Publications/JASIS/Best_Jasist/1976pricejasistarticle.pdf

Rodriguez MA, Bollen J, van de Sompel H (2007) A practical ontology for the large-scale modeling of scholarly artifacts and their usage. In: Larson R (ed) JCDL 2007: 7th ACM/IEEE joint conference on digital libraries 2007: Building and sustaining the digital environment, June 18-23, 2007, Vancouver, British Columbia, Canada. ACM, New York, NY, pp 278–287 (DOI: 10.1145/1255175.1255229), also available as arXiv preprint arXiv:0708.1150

Small HG (1978) Cited documents as concept symbols. Soc Stud Sci 8(3):327–340 (DOI: 10.1177/030631277800800305, stable JSTOR URL: http://www.jstor.org/stable/284908)

Wagner CS (2008) The new invisible college: Science for development. Brookings Institution Press, Washington, DC

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. Nature 393(6684):440–442 (DOI: 10.1038/30918)

Whitley R (1984) The intellectual and social organization of the sciences. Oxford University Press, Oxford

Wojick DE, Warnick WL, Carroll BC, Crowe J (2006) The digital road to scientific knowledge diffusion: A faster, better way to scientific progress? D-Lib Magazine 12(6) (DOI: 10.1045/june2006-wojick)

Zhang H, Qiu B, Ivanova K, Giles CL, Foley HC, Yen J (2010) Locality and attachedness-based temporal social network growth dynamics analysis: A case study of evolving nanotechnology scientific collaboration networks. J Am Soc Inf Sci Technol 61(5):964–977 (DOI: 10.1002/asi.21225)

Zipf GK (1949) Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley, Cambridge, MA

Zuccala A (2006) Modeling the invisible college. J Am Soc Inf Sci Technol 57(2):152–168 (DOI: 10.1002/asi.20256)

Glossary

Agent In the context of an agent-based model, an agent is an individual that is capable of autonomous behavior. It usually has a well-defined internal state and is situated in an environment with which it can interact. That environment usually includes other agents and other targets of interaction.

Base knowledge Facts and ideas that are more or less widely known within a *specialty*. These can correspond to widely accepted ideas and theories, techniques, and empirical facts, but can also correspond to controversies or conflicting ideas. Base knowledge is most often referred to by citing the *documents* in which those facts or ideas were either first or most prominently elucidated. Cited documents, or references, are thus used as symbols for base knowledge.

Citation Citation is a term that can be easily misunderstood. It is used in two different senses by different groups of researchers. In the biomedical and social science literatures, “citation” typically refers to a *document*, or a *node* in the document network. For example, a MEDLINE citation refers to the bibliographic record of a document. By contrast, in bibliometrics and network science, "citation" refers to the directed *link* between one document and another; it refers to the citation of one document by another document. Citation counts thus accrue to cited documents. In citation analysis, one speaks of a document having been cited *n* times, or having *n* citations. In this work we use the bibliometric definition of "citation" exclusively.

Clustering The process of assigning a set of elements into groups, where the elements in a group are similar to each other in some sense (e.g., according to selected properties of units). In the three network types listed here, *researchers*, *documents*, and *terms* can each be clustered into groups based on similarities in those elements. Although the individual elements of a network are the basic units, clusters are often the unit of analysis that is reported. Clustering is often used to approximate the composition of conceptual aggregations. For example, authors can be clustered to approximate the memberships of different *invisible colleges*, documents can be clustered to approximate the outputs of *research communities* or *specialties*, and terms can be clustered to form broader *topic* spaces.

Collaboration Collaboration is an active process where two or more researchers and/or institutions work together on something of common interest. Co-authorship of a document is thought of as a direct *indicator* of collaboration.

Communication Communication in science can happen on a variety of levels, both formal and informal. It is the mode by which an invisible college operates, and can include everything from the most formal *collaboration* (co-authorship, which is relatively easy to measure), to the transmission of ideas through the reading and citing of articles (measurable), to informal discourse on scientific topics via face-to-face, phone, or email conversations (far less measurable).

Discipline An academic or scientific discipline (or field) is an established body of knowledge with similar cognitive content. This establishment, while fundamentally cognitive, is most clearly evidenced in the existence of interconnected social and institutional structures (or networks), such as discipline-specific university departments or institutes where research is performed and instruction takes place, as well as in discipline-specific academic *journals*, organizations, societies and meetings. Disciplines fulfill a number of roles: they specify the objects that can be studied, provide methods, train and certify practitioners, manufacture discourse, provide jobs, secure funding, and generate prestige. Some of the traits a discipline should have are: university departments and institutes, specialized scientific societies, specialized journals, textbooks, a specific domain of objects studied from a specific perspective, methods for the production and analysis of data, means of presentation using specific terminology as a conceptual framework, and forms of communication. In science modeling, a discipline is most often defined as a set of journals, or as the papers published in a set of journals. Some people refer to a discipline as a large set of papers around a particular *field* of study, without regard to a particular set of journals. We prefer to call this type of aggregation a field rather than a discipline.

Document For science and bibliometrics studies, scientific articles are usually the basic independent record in the project database. Documents can include various article types, including journal articles, review papers, conference papers, etc. If extended beyond the scientific realm, documents can include gray literature, government reports, patents, and even the proposals associated with funded research.

Element Individual vertices or nodes.

Funding Monetary inputs into the science system. These can come in the form of grants, contracts, investments (e.g., venture capital), or direct R&D monies within an institution.

Indicator "Science indicators are measures of changes in aspects of sciences"
AQ2 (Elkana, Lederberg, Merton, Thackray, & Zuckerman 1978).

Institution In the context of science modeling, an institution is an organization that creates knowledge, typically through the mechanism of an author publishing an article. In a practical sense, institution names are typically listed with author or inventor names in *documents*. Institutions can also include funding agencies.

Invisible college The most recent definition of invisible college comes from (Zuccala 2006): "An invisible college is a set of interacting scholars or scientists [*researchers*] who share similar research interests concerning a subject *specialty*, who often produce publications [*documents*] relevant to this subject and who *communicate* both formally and informally with one another to work towards important goals in the subject, even though they may belong to geographically distant research affiliates."

Journal A publication medium in which a selection of scientific articles (*documents*) on a particular topic or set of topics is published, typically in a series of issues. A journal can appear in print or electronic form or both. Most journals that are considered as the prime publication outlets by researchers are peer-reviewed, meaning that other researchers review submitted manuscripts and recommend (or not) their publication.

Knowledge diffusion The process by which science knowledge is spread (Wojick et al. 2006).

Knowledge validation Peer review and replication.

Network A network is a set of vertices (or **nodes**) that represent the units, and a set of lines (or **links**) that describe the relationship between those elements. Networks are often represented visually by graphs using node/link diagrams. Many different networks can be created from bibliographic data – for example, a social network showing the relationships between people (*researchers*), a knowledge network showing relationships between *documents*, or a descriptor network that shows relationships between *terms*.

Network metric A variety of metrics are used to characterize properties of networks. These include edge count distributions (known as degree, in-degree, or out-degree), path lengths, clustering coefficients, centralities of various types, etc.

Researcher As a broad definition, a researcher is a person who performs research. In terms of modeling science, a researcher must not only perform research, but must also publish that research. For the purpose of modeling of science and technology, we can expand that definition to include authors who publish, inventors who apply for patents, and investigators who apply for and receive funding through grant proposals.

Research community Many years ago, sociologists, specifically Kuhn (1962) and Merton (1973), suggested that **researchers** organize themselves into relatively small socio-cognitive groups – on the order of 10 people – working on common problems. Although the word "community" implies a group of people, the output of a single such group can be thought of as a research community. A typical community will publish around 10–15 articles (*documents*) per year, assuming the authors each publish 1–2 articles annually on the problem focused on by the community.

Research front The working definition of a research front according to Thomson's ScienceWatch is that of a co-citation cluster of highly cited articles, limited to the most recent 5 years. A more general definition might be "a specialty's current literature" or "the most recent development of a specialty" without regard to being highly cited or not.

Research specialty A research specialty (or field) is usually defined at a higher level of aggregation than a **research community**, and can be thought of (more or less) as the documents published by an invisible college. A research specialty can be comprised of many research communities and is comprised of, on average, hundreds of articles per year. Lucio-Arias and Leydesdorff (2009) write that "a research specialty can be operationalized as an evolving set of related documents. Each publication can be expected to contribute to the further development of the specialty at the research front." Research specialty is often considered to be the largest homogeneous unit of science, in that each specialty has its own set of problems, a core of researchers, shared knowledge, a vocabulary, and literature.

Team A small group of **researchers** who tend to work together on a particular topic or set of topics. Members of research teams are strongly connected – that is, each team member knows and interacts with, and often co-authors with, the other team members. Teams are typically low-level groups that cannot be further subdivided.

Term A single- or multiple-word phrase. Terms can be generated in different ways. For instance, they can be chosen from a standardized set of terms (e.g., a thesaurus like MeSH) by an author, indexer, or editor; or they can be extracted from a document, title, or abstract using automated means.

Topic A topic can be an area of interest or the focus of an article or *document*. The notion of topic includes both a main idea and supporting details. Thus, a topic is much broader than a single *term*.

Unit Element type (e.g., author, article, journal, etc.).

AUTHOR QUERIES

AQ1. First author has been considered as corresponding author. Please check.
AQ2. Please cite this reference in text.

Chapter 2
Mathematical Approaches to Modeling Science from an Algorithmic-Historiography Perspective

Diana Lucio-Arias and Andrea Scharnhorst

2.1 A Narrative of the History of Mathematical Models of Science

The accumulative nature of knowledge requires systematic ways to comprehend and make sense of what we know. In the case of scientific knowledge, this requirement is enhanced by the importance given to science as a driver of social and economic progress. The persistent interest in a "science of science" or a "social studies of science" is a consequence of the reflexive endeavor to comprehend and assimilate science and the growth of scientific knowledge – perhaps together with policy intentions to design evaluation and stimulus mechanisms.

This interest has led to significant efforts to define and refine ways of modeling, representing, and understanding science in the scientific community – efforts unrestricted to single disciplines or intellectual traditions. Reflection upon knowledge production co-evolves with knowledge production itself. It reaches from early philosophy to the arts, encompassing attempts to order knowledge. One famous example of how to order knowledge is the arbor scientiae of the philosopher Raimudus Lullus (1232–1316) (Dominguez Reboiras et al. 2002).

At the same time, in our modern understanding, the old symbol of the tree also encompasses the idea of evolution. To characterize the evolution of the science system (natural sciences, social sciences, humanities, and arts), its growth and differentiation, mathematical models are one possible scientific method. This book reviews

D. Lucio-Arias (✉)
AQ1 Colombian Observatory of Science and Technology, carrera 15 no 37-59, Bogotá, D.C, Colombia
e-mail: dlucio@ocyt.org.co

A. Scharnhorst
Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences,
Anna van Saksenlaan 10, 2593 HT Den Haag, The Netherlands
e-mail: andrea.scharnhorst@dans.knaw.nl

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_2,

the transfer of models belonging to different branches of an imagined "(sub)tree of mathematics" to scientometrics. Mathematical models in scientometrics are developed to understand better the structure and evolution of the imagined whole tree of knowledge, and so the circle closes. In this chapter, the metaphor of the tree reoccurs once more in the method used to depict the history of mathematical modeling of the sciences. Treelike structures are the core of the historiographic method where, constructed from citations of key papers, they illustrate the evolution of knowledge.

Mathematical models of the sciences do not stand alone in our modern day but stem from formulations made earlier in time. Mathematics has penetrated almost all other scientific disciplines. We not only know mathematical physics and mathematical biology, but also mathematical economics, mathematical sociology, mathematical psychology and mathematical finance.[1] Although there is no field of "mathematical science studies," the emergence of quantitative studies of science – bibliometrics, scientometrics, informetrics – came along naturally together with mathematical approaches. Not surprisingly, methods of statistics are well established in scientometrics (Egghe and Rousseau 1990). However, applications of mathematical models to the dynamics of the science system form relatively singular and isolated events. This observation, together with an increasing need for modeling dynamic processes in science, was not only the trigger for this book, but also the starting point for this chapter.

We can attempt to categorize mathematical models of science according to the phenomena they try to explain and the epistemic approaches they follow. Phenomena include: growth and distribution of expenditures for education and research across countries and fields; number of PhD's in different fields; growth of the number of publications; formation of and competition between scientific fields; citation structures; and different productivity patterns among researchers from different disciplines, taking into account age and gender. Epistemic approaches differ according to their perspective (which can be micro or macro), their basic elements, their units of analysis, and how major dynamic mechanisms of the system under study are identified. Scientific methods are part of the epistemics, so models of science can differ by their use of mathematical technique and mathematical language (see Börner et al. in Chap. 1). Concerning mathematical approaches applied to the sciences as an object, we observe a mixture between new mathematical techniques available and newly emerging scientific fields.

In Fig. 2.1, we try to sketch the appearance and diffusion of some mathematical models of science. This sketch is based on the insights of one author who did

---

[1] The appearance of separate subject classifications for these subfields or specialization in the Mathematics Subject Classification (MSC) – a system used to categorize items covered by the two reviewing databases, Mathematical Reviews (MR) and Zentralblatt MATH (Zbl) – can indicate the consolidation of mathematical approaches in these fields. According to the MSC2010, mathematical economics encompasses 37 subclasses, mathematical sociology 6, mathematical psychology 5, and mathematical finance 9 (see http://www.ams.org/mathscinet/msc/msc2010.html).

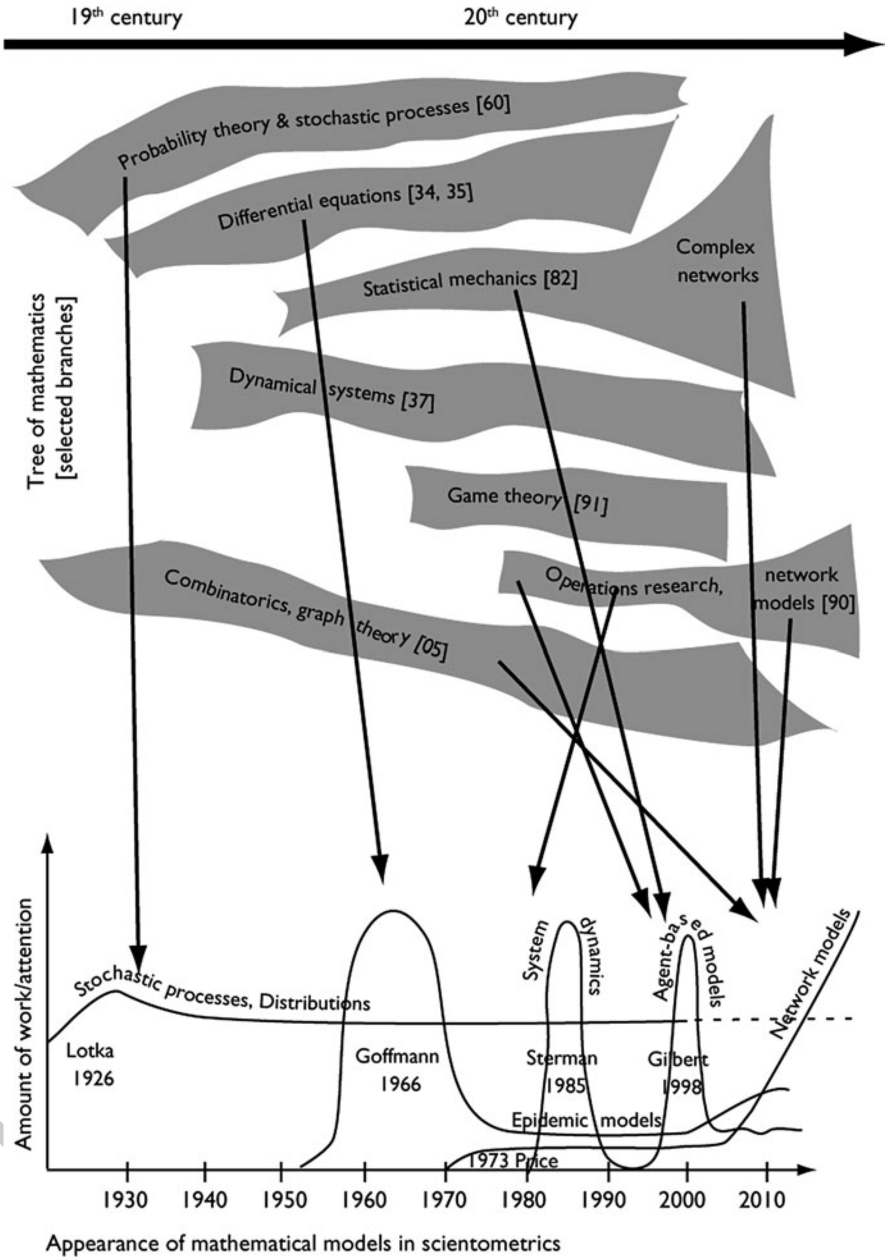


Fig. 2.1 Branches of mathematics and appearance of mathematical models in scientometrics

her PhD in this area in 1988 and kept publishing in the field (Scharnhorst 1988; Bruckner et al. 1990). In the upper part of Fig. 2.1, branches of mathematics are selected (labeled according to the Mathematics Subject Classification) according to their relevance for models of science. Of course, inside mathematics, these branches overlap and form a fabric (Boyack and Klavans 2009), or turbulent, reacting-diffusing fluids, rather than a static tree with separable branches. The lower part of Fig. 2.1 depicts growth curves of certain models of science. However, there is no linear causality between a certain progress in mathematics and its possible application to the science system, even if we indicate relations by arrows as in Fig. 2.1. Few models enter the field of scientometrics via biology, psychology, economy, or physics. Last but not least, it all depends if researchers are intrigued enough by the problem to model mathematically the sciences as a cognitive and social system.

For the time being, we would like to stick to such a narrative that combines epistemic streams running across different disciplines with the first occurrence of certain types of models applied to science as a system. In the main part of the paper, we search for empirical evidence supporting or contradicting this historical narrative.

We state that in parallel with the emergence and spreading of "approaches and techniques" (for example, stochastic distributions at the end of the nineteenth century; the emergence of system science and operations research; the paradigmatic change in physics towards irreversible, dissipative and complex processes; and the rise of rule-base agent modeling, to name only a few), researchers – most of the time also pioneers in developing these methods – were curious also to apply them to an environment in which they felt at home: the academic system.

For instance, Lotka described the skewed distribution of the productivity of scientists (Lotka 1926) as part of his more general approach to apply methods of (statistical) physics to evolution in nature as well as society (Lotka 1911). Sterman's system-dynamics model of Kuhn's scientific revolution (Sterman 1985) is embedded in his overall work on complex social systems, part of the emergence of system dynamics as a specific mathematical systems theory (Sterman 1992), and just another exemplification of feedback loops and complex correlations between dynamic micromechanisms. Goffman modeled the diffusion of ideas similarly to the spreading of diseases, and other researchers (Nowakowska, Kochen, Yablonsky, Bruckner et al.) compared the emergence of scientific fields to the evolution of biological species. They all made use of differential equations and master equations at the moment non-linear differential equations became very popular ways to describe the dynamics of complex systems (Nicolis and Prigogine 1977). Gilbert's agent-based model of science (1997) marks the entry and spread of rule-based modeling into mathematical and computational sociology (Epstein and Axtell 1996), for which Gilbert also did pioneering work (Gilbert and Troitzsch 2005). Furthermore, the interest of Gilbert was also obviously triggered by his earlier work on the history and sociology of science (Gilbert and Mulkay 1984).

But not in all cases do we find a strict temporal correlation between the establishment of the mathematical method and its testing out for the science system as one specific social system. In the case of game theory, developed in the 1930s and 1940s (see von Neumann and Morgenstern 1944), only now is the method tested upon science itself (see Hanauske in Chap. 5).

Moreover, there are differences in the way the scientific community has embraced these pioneering approaches. Lotka's law is known today as a classic law in scientometrics. Stochastic processes, which can explain also Lotka's law, have been present almost the whole time (e.g., Glänzel and Schubert 1995; Van Raan 2006; Egghe 2005. However, Lotka's general framework of a physics of evolution applicable to processes in nature and society did not travel. Even more, his famous systems of non-linear differential equations (Lotka–Volterra equations), applied extensively in mathematical biology (Lotka 1925), did not travel, at least not through Lotka's own initiation. Although Goffman's epidemic model belongs to the same type of models, the link to Lotka–Volterra equations has been made explicit only in the 1980s. After seeing a first rush in the 1960s, 1970s and 1980s, epidemic models themselves only reappeared in the context of epidemic processes on networks, together with the emergence of a cross-disciplinary network science (2005), from 2000 onwards. In the same context of the revival of networks, other early network models like Price's gain a second period of attention. In contrast, applications of agent-based models and system-dynamics models remain rare occurrences. Yet, agent-based models – outside of scientometrics and independent from it – have been embraced by computational philosophy, which uses concepts and mathematical approaches for epistemic spaces and dynamics quite similar to those used in scientometrics ((Weisberg and Muldoon 2009) see Payette in Chap. 4). All in all, the impression emerges that mathematical models applied to science come in waves, remain relatively independent from each other, and form more an ephemeral than a persistent thread in scientometrics (Fig. 2.1).

This is quite interesting. Why, unlike other sciences, does the modeling of science dynamics appear as a process of eternal beginning, and why does it still lack a coherent theoretical framework? Can we find facts for such an impression now turned into a hypothesis? Can bibliometrics confirm that we indeed are faced today with modeling approaches to science that are scattered, while older approaches might have been obliterated or forgotten with time? Can historiographic analysis also reveal some of the causes for such a situation?

The purpose of this chapter is to counter an individual account of science history with a bibliometric study. We present a historiography of mathematical models and approaches to science. This will give the opportunity to reveal the cognitive history of the models. What might seem unrelated today might share a cognitive or disciplinary memory or might stem from significant older papers that had citation relations between them. We follow this section with a description of the method of algorithmic historiography to reveal scientific developments. This method is later used to (a) delineate the cognitive historiography of today's mathematical approaches to science and (b) illustrate approaches to science constituting a lasting thread that may have been forgotten or obliterated by new models.

2.2 The Use of Bibliometrics in Science History – Algorithmic Historiography

Publishing as a means of communicating, corroborating, or refuting scientific findings is a crucial operation for the development of scientific knowledge (Lucio-Arias and Leydesdorff 2009). For this reason, citation practices have also become established in this discursive construction of scientific knowledge (Wouters 1999). Early in the invention of citation indexing, which was primarily aimed at advancing information retrieval, Garfield proposed to use these databases to reconstruct the history of scientific ideas (Garfield et al. 1964). The bibliographic information contained in a collection of published articles and their references makes historical reconstruction through citations a collective and social enterprise (ibid.). However, one has to keep in mind that looking at citations represents a specific empirical method. Both bibliometrics and scientometrics have known a long and continuing debate over the meaning of citations in knowledge production, dissemination, and reconstruction (De Bellis 2009). Recently, it has been observed that "it remains a question what actually bibliometrics can add to science history based on text analysis and eye witness accounts" (Scharnhorst and Garfield 2010). The method of algorithmic historiography as applied in the following is therefore used as one possible empirical method to test some of the hypotheses presented in the previous section, and the results make explicit the limitations of this method.

The notion of algorithmic historiography is supported by the introduction of HistCite$^{\mathrm{TM}}$ as a bibliometric tool that aids the process of uncovering transmissions of knowledge that lead to scientific breakthroughs (Pudovkin and Garfield 2002). It relies on citation data to describe historically scientific fields, specialties, and breakthroughs (Garfield 1979). The software creates a mini-citation matrix for any set of documents retrieved from the ISI Web of Science, facilitating historical reconstructions based on a literary simplification of science (Garfield et al. 2003b,a, 2005). Depending on the seed nodes selected to start the citation, mining the method can be applied to a scientific field or a journal, the oeuvre of a scholar, or an individual paper (Scharnhorst and Garfield 2010).

The method of utilizing the textual footprint of scientific discoveries and breakthroughs to reconstruct their history has been employed in scientometrics. Citations might be considered as the memory carriers of the system, and their use as nodes in network-like historiographs can be further enhanced by using algorithms from network and information theory (Lucio-Arias and Leydesdorff 2008). Even though this approach is used to a lesser extent by philosophers and historians of science, the algorithmic approach to historical reconstruction enables us to include more variety in the perspective than a reconstruction based on dispersed narratives (Kranakis and Leydesdorff 1989). This approach, labeled scientometric historiography, relies on citation networks to build descriptive reconstructions of history, assuming that these networks reflect a transmission or flow of ideas between papers.

Possible biases caused by the use of citations for empirical reconstructions might include the overestimation of contributions from elite scientists (MacRoberts and MacRoberts 1987, 1989), negative or critical citations, or the perfunctory acknowledgement of earlier work. Nevertheless, different studies have agreed that around 70% of the references used in a scientific paper correspond to criteria of scientific relevance (Vinkler 1996; Krampen et al. 2007). In other words, 70% of citations respond to the normative theory of citing (Cronin 1984), which justifies the value of citation analysis for historical reconstruction of scientific fields. We use the main-path algorithm from social network analysis to identify those central documents in the citation networks. Specifically, we use the Search Path Link Count available in Pajek which accounts for the number of all possible search paths through the network emanating from an origin (Hummon and Doreian 1989; Batagelj 2003). These main paths have been acknowledged to identify documents that build on previous work, while acting as authorities for later works (Yin et al. 2006). These documents can be expected to be associated with thematic or methodological transitions in the development of a topic (Carley et al. 1993) and are significant for writing the history of science (Hummon and Doreian 1989).

In the following sections, we use two different approaches to chronological networks of citations. Citations allow us to study the diffusion of ideas among documents. But citations can also be understood in the process of codifying scientific knowledge. They link older texts to today's scientific knowledge while providing information about the cognitive position of scientific knowledge claims, which through citations and references get contextualized in scientific repertoires and trajectories. Citations give disciplinary context to publications. We will take both of these perspectives into account in the following sections. In the first part of the results section, we will present the bibliographic history of mathematical models used today to study science. We expect to encounter well-known pioneers like the models mentioned throughout the book, but we will also encounter lesser-known models that may have been obliterated or forgotten over time. We will show how different threads are codified in relation to different "classical" or seminal approaches to mathematical models of science. The second reading given in the results section corresponds to the trajectories constructed from the diffusion of seminal approaches to science modeling. We reconstruct the diffusion of the ideas introduced by Alfred J. Lotka, Derek de Solla Price, and William Goffman based on citation analysis.

2.3 Data Selection and Analysis Design

In this chapter, we use bibliometrics to study and follow the implementation of mathematical models for science. The purpose will be to uncover different characteristics of the process of codifying mathematical models that have been published in the last 5 years in selected journals of Library and Information Science. In this section, we look at the knowledge base of this set of papers to determine their cohesiveness. The method of using mathematics to model the structure and behavior of science presents scattered trajectories that could respond to the lack of a unifying theory or intellectual base. In a later section, some of the models that appear in chapters of this book will be presented from the perspective of their

Table 2.1 Statistics of the search: present to past

Journal	Documents	Inside citations	Total citations
JASIST	50	39	416
Scientometrics	47	44	271
IP&M	20	5	63
J. Informetrics	20	13	53
Total documents		137	

Source: ISI Web of Science, query May 25, 2010, HistCite$^{\mathrm{TM}}$.

diffusion trajectories. This will emphasize possible recombinations, cognitive links, or disciplinary shifts that affect the appropriation of the models in the scientific community. In this specific section, the diffusion trajectories are detailed in relation to the characteristics of the models presented in the introductory chapter of this book.

All our analyses are based on retrievals from the Thomson Reuters Web of Science, which can easily be read by the HistCite$^{\mathrm{TM}}$ software.

For the cognitive history of contemporary papers using (or referring to) mathematical models of science (Present to past analysis – Sect. 2.4.1), we selected four major journals in ISI's subject category of Library and Information Science. The selection of the journals was determined by their popularity inside the community of the information sciences. For retrieving documents using mathematical approaches to science, we first used a topical search in the ISI Web of Science[2] that retrieved 2,876 documents. However, we encountered the problem that the majority of them were not in line with the purpose of our study. For this reason, we decided to download all documents published in Scientometrics, Journal of the American Society for Information Science and Technology, Journal of Informetrics, and Information Processing and Management in the period considered. We made a manual selection based on the titles, abstracts, and full text (when necessary) of those documents that used mathematical approaches (ideally models) to explain science. The drawback of this last approach is that there are various mathematical models in existence. There is also an ambiguity in the use of the word "model" and even "mathematical model." Many of the documents selected claimed to be modeling approaches but failed to have all the specifications necessary to be considered as such. Table 2.1 gives an overview of the number of retrieved documents per journal, as well as the citations inside the retrieved set of documents (inside citations) and in the whole web of science (total citations).[3] Table 2.1 also presents a summary of the volume of papers selected according to the sample of journals taken. The whole set of 137 documents selected as referring to mathematical models of science for 2005–2010 is available at the end of this chapter in Appendix 1.

The software HistCite$^{\mathrm{TM}}$ was used to build the inner-citation matrix of these documents to illustrate their cognitive relatedness. Because they might be related in a citation window larger than the years considered, the set was expanded to include the most highly cited documents inside the set.

---

[2] Query used: ts = (model* same (science or scientific or knowledge)).

[3] For comparable analysis, the whole data set can be requested from the authors.

For the second part of the analysis, the diffusion trajectories of three different models were chosen according to their relevance and impact in scientometric studies. We chose Lotka's law, Goffman's epidemic model, and Price's network model. The three models differ in character. Lotka's law is a statistical description (a descriptive model) of certain structures in science. Goffman's model departs from assumptions of basic mechanisms of science on a micro level to reveal structures on a macro level due to the dynamics imposed. It can be used for description as well as for prediction. Price's network model is a conceptual one that reflects upon possible disciplinary meanings that emerge from the network structures formed by citation relations between papers. It is empirically verified and exemplifies phenomena such as obliteration, the relation between references and citations, and the emergence of research fronts. However, there is only a small step between descriptive models and predictive models. Distributions, as in the case of Lotka's law, have been explained from stochastic processes. Price has himself later proposed mathematical models for the micromechanisms behind some of the features he explores in his "Network" paper (Price 1976). The popularity of Lotka's law as one of the few basic laws of science and the fact that it operates at the border between descriptive and predictive models were the reasons we included Lotka's law in our selection. In the case of Price's network model, we chose an example of a comprehensive and classical description of a basic pattern in scientific communication that has inspired

Table 2.2 Seed documents

Model	Seed documents	# cites (papers considered)	Citation window (in years)
Lotka–Volterra model	Lotka, A.J. (1926). The frequency distribution of scientific productivity, J. Wash. Acad. Sci., 16: 317	612	1939–2010
Price network model	Price, D.J.D. (1965). Networks of scientific papers. The pattern of bibliographic references indicates the nature of the scientific front, Science, 149 (3683): 510-515	497	1978–2010
Goffman epidemic model	Goffman, W. (1966). Mathematical Approach to Spread of Scientific Ideas – History of Mast Cell Research, Nature, 212 (5061): 449 Goffman, W., & Newill, V.A. (1964). Generalization of Epidemic Theory: An Application to the Transmission of Ideas, Nature 204: 225.	73	1975–2010

Source: ISI Web of Science, query May 25, 2010.

many other reflections, some of them mathematical. We explain each model at the beginning of the corresponding results section.

Table 2.2 depicts the documents that were used as seed documents for these models. It shows the amount of times the chosen seed documents were cited and the publication years of those citing documents. All documents citing these seeding documents were downloaded and analyzed according to their modeling characteristics.

The downloaded citing documents were content analyzed to identify the purpose of the paper (if it was a mathematical approach, an application or refutation of informetric laws with empirical evidence, an evaluation or assessment exercise in a specific context, etc.).

2.4 Results

2.4.1 The Current Presence of Mathematical Modeling in Library and Information Science – Following Traces from the Present to the Past

To analyze the intellectual base of the papers that are currently applying mathematical models to study science, we started from our sample database (Table 2.1), which consists of 137 documents published in leading journals in ISI's subject category of Library and Information Science from 2005 to 2010. These papers were taken as seeds for a HistCite^{TM} analysis with the purpose of tracing the citation relations inside the set. The resulting historiograph (Fig. 2.2) depicts documents as nodes, where the size of the node represents the amount of citations it gets inside the considered set (outside citations are not taken into account). The arrow represents a citation relation. We start from the current papers, dig into their bibliographies and look for cross-connections. We also try to see how persistent models are, and which mathematical models we encounter.

Figure 2.2 shows the citation diagram for the current mathematical approaches to science. The number of the nodes corresponds to the numbers of the 137 documents in the first appendix. Most of the nodes are related to stochastic processes in informetric data.

Already, one sees that the documents dealing with mathematical models belong to different, isolated threads. We present a zoom of four of them in the subsequent

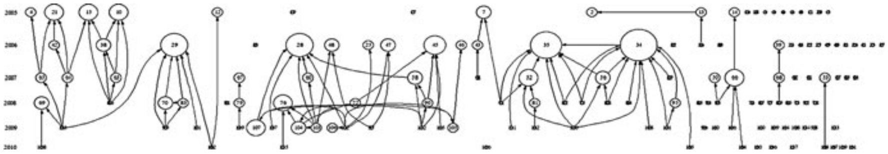


Fig. 2.2 HistCite^{TM} output of papers using mathematical approaches to understand the science system – overview

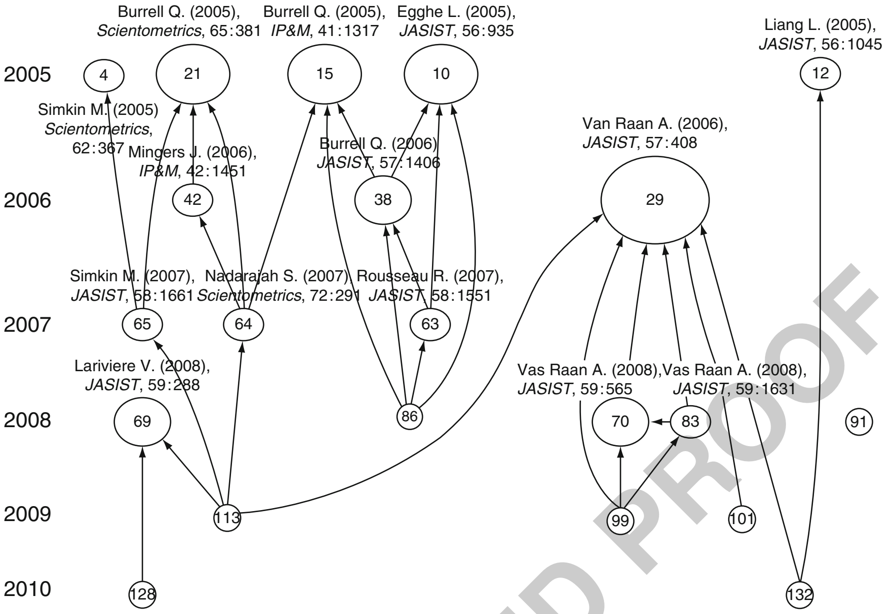


Fig. 2.3 First thread in the current HistCite$^{\text{TM}}$ graph of mathematical model papers

figures and label the nodes that are cited inside the set with their bibliographic references.

In the first group, from left, we find a paper by Van Raan (29) about statistical properties of indicators. Some of the papers in our set emphasize modeling and explaining through mathematical formulations citing behavior and growth (e.g., Nodes 4, 10, 15, 21, 38, 42, 63, 64, 65, 83 and 108).

As we move in Fig. 2.2 from left to right (or from Figs. 2.3 to 2.6), more sophistication is added to the approaches, going from explanations and refinements based on the Hirsch index, to model impact and relevance of authors, to research group behavior (e.g., Nodes 29, 70, 83). However, most of the papers explain the static structure of science. In the last few years, the efforts that have been undertaken to explain growth in the system of science seem unrelated to the rest of the papers (e.g., Nodes 2, 13, 70, 76, 96).

In the second group, we find papers about network algorithms and approaches to mapping science – particularly, old and new approaches (Small 48, Börner 46, Klavans 47) and Chen's citespace software (28). This thread interestingly binds mapping and network approaches with predictive models on epidemics of idea spreading (Bettencourt 76) and the peer review process (Bornmann 67). (A list of all papers is given in Appendix 3.) All the nodes for the year 2009 correspond to the "Science of Science" special issue of the Journal of Informetrics.

A third group entails a paper about statistical features of the Hirsh-index, the newest challenge to bibliometric rankings (e.g., Nodes 34, 35, 56).

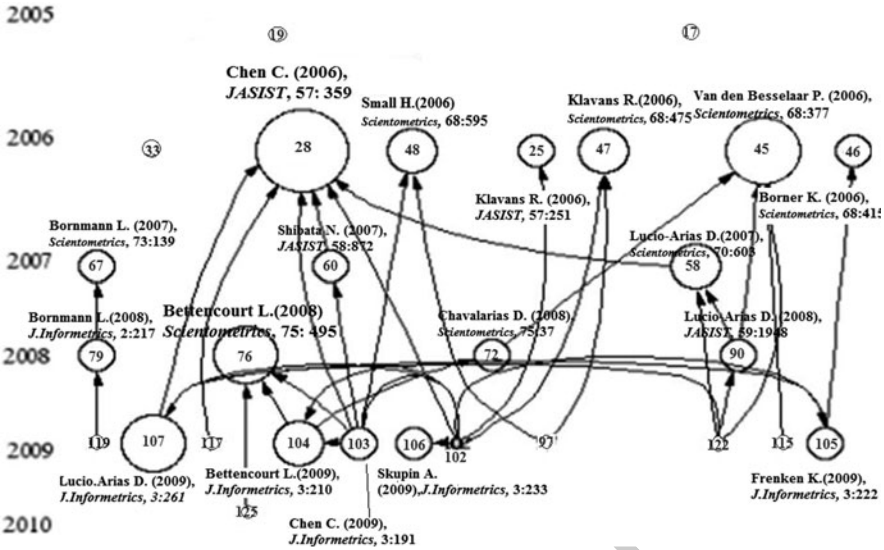


Fig. 2.4 Second thread in the current HistCite$^{\text{TM}}$ graph of mathematical model papers

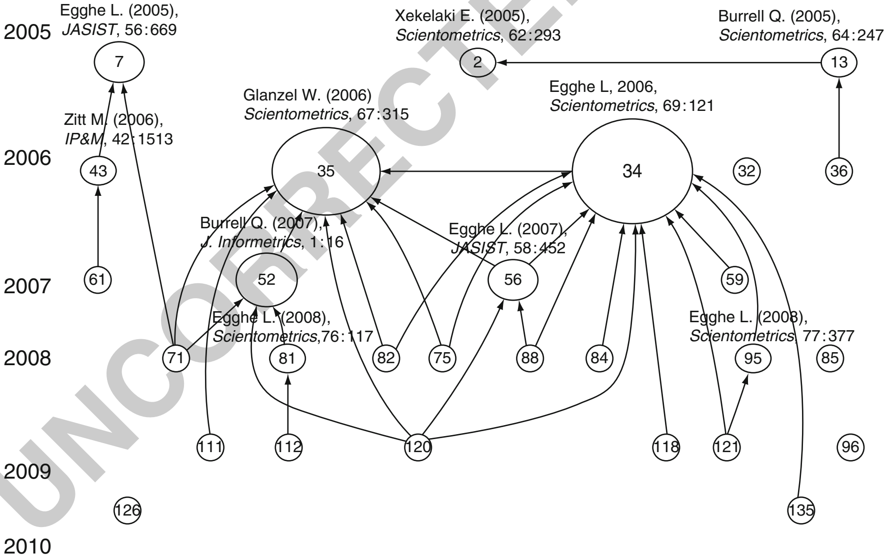


Fig. 2.5 Third thread in the current HistCite$^{\text{TM}}$ graph of mathematical model papers

Comparing our analysis of the different threads with Fig. 2.7, one can see that although many of the documents treat similar issues (especially stochastic behavior), there is no clear relation between them. For instance, Node 76 (in Fig. 2.4) represents

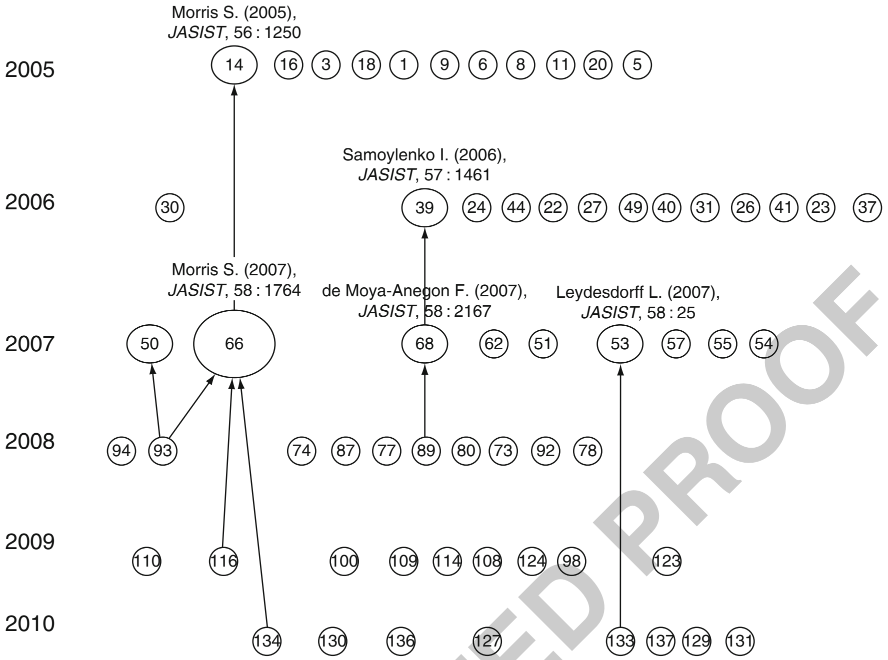


Fig. 2.6 Fourth thread in the current HistCite$^{\mathrm{TM}}$ graph of mathematical model papers

the paper by Bettencourt, Kaiser, Kaur, Castillo-Chávez & Wojick from 2008 that reuses the model of epidemic approaches for the transmission of ideas; as can be seen in the historiograph, this node does not have any citation relation with the other papers in the set.

Strikingly, the bibliometric analysis seems indeed to confirm the historic narrative. Mathematical models of the sciences are divided into different branches and exist largely in isolation, as can be seen by the occurrence of many single points at the right side of both Figs. 2.2 and 2.7.

The isolation of the sets might respond to functional differentiation that results from the growth in scientific publications, and that allows scientists to reduce the levels of complexity in different disciplines (Lucio-Arias and Leydesdorff 2009). This means that the apparent isolation between sets might be reduced when looking at the bibliographic antecedents of these models. In Fig. 2.7, the most cited documents outside the set of the 137 documents selected for treating science with mathematical models and approaches were incorporated to construct a new historiograph.

From Fig. 2.7, it can be deduced that, even if different papers are not closely related to other contemporary approaches, they seem to have a common cognitive historiography, and there is a consensus on classical or seminal approaches to current modeling exercises to understand the sciences. In Fig. 2.7, the main path

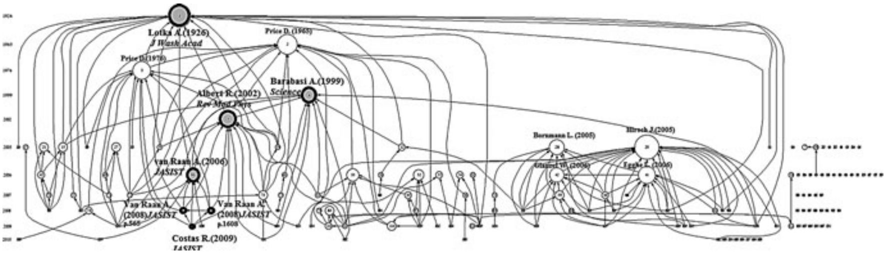


Fig. 2.7 HistCite^{TM} output of papers using mathematical approaches to understand the science system, enhanced with their cognitive history

of the set is highlighted in gray. Lotka's seminal paper, which originated Lotka's law on scientific productivity based on the skewed distributions of authors, is the starting point; due to the interdisciplinary nature of the paper, the next two documents highlighted in the main path – Barabási (1999) and Albert (2002) – are also foreign to the field of Library and Information Science and, more specifically, to scientometrics. These papers deal with networks as random graphs from a physics perspective; the next nodes in the main path (36, 77 and 90 – Van Raan (2006, 2008a, 2008b) reflect the discourse about the importance of impact upon research groups and individuals. Interestingly, from this wider perspective, statistical physics and complex networks, as well as rankings and indicators, seem to be interwoven into one network of exchange of ideas.

AQ2

The scattered impression depicted in Figs. 2.2–2.6 reflects the sparse relatedness of mathematical approaches inside of Library and Information Science. It can also be interpreted as a lack of consolidation around mathematical methods and as competition between different threads of mathematical modeling that are related in principle but divided in practice. Figure 2.7 shows that when overlooking larger parts of the scientific landscape, these isolated branches or points are interconnected. One could say that the generic and universal character of mathematical approaches that can act as bridging and transporting structures of knowledge diffusion is more visible in Fig. 2.7. In any case, the comparison of Figs. 2.2 and 2.7 shows the relevance of the selection of the seed nodes. It also shows the restriction of a too inner-field perspective. The position of mathematical modeling in scientometrics cannot be fully understood from the field's perspective only. We need to look at the tension of evolution inside of one field and among different fields. "Neighboring fields"[4] of Library and Information Sciences might be seen as a relative constant and as a neglected environment if it concerns threads inside of LIS that are mature. For a rather marginal topic such as dynamic models of science, they gain importance as a source of ideas travelling into LIS.

For Fig. 2.7, the set of 137 documents dealing with mathematical approximations to science from the perspective of Library and Information Science was studied;

---

[4] Independently how we define neighborhood here.

included in the set were the most highly cited documents (144 documents in total). While the recent documents could be considered the research front of the field, the highly cited ones can be considered the intellectual base (Chen 2006). The main path has been acknowledged in scientometric studies to represent the backbone of a journal or a field (Hummon and Doreian 1989; Carley et al. 1993). Nevertheless, the main path depicted in Fig. 2.7, although highlighting important documents in the topic of mathematical models of science, cannot be taken as the main achievements of the field. The reason is that the set does not represent a cohesive specialty or discipline.

We used bibliographic coupling of authors to measure cognitive cohesiveness in terms of similarities between reference lists in the set of papers. This coupling technique uses author names as variables and the references as cases. To correct for productive authors with many papers, cosine normalization is applied. Figure 2.8 illustrates the results for 187 authors publishing mathematical models of science.

While Figs. 2.2–2.7 illustrate the citation network as a chronological network of citation where documents are organized according to their publishing year and their bibliographic antecedents and descendents, the coupling in Fig. 2.8 corresponds to authors based on the similarities of the referenced works in their papers. It supports the suggestion of Fig. 2.8 of a common cognitive history in these approaches to modeling science.

2.4.2 The History of Mathematical Modeling of the Science System – Following Traces from the Past to the Present

2.4.2.1 Lotka, Goffman, Price: Overall Growth and Diffusion of Reception

In this section, we present the diffusion trajectories of three specific models: Lotka's law (as discussed in Chap. 3 of this book), Goffman's epidemic model (see also Chap. 3), and the network model introduced by Derek de Solla Price (addressed also in Chap. 7 of this book, Fortunato et al.). Even though the three models remain very relevant in the information sciences, their impact measured in terms of citations varies (see Fig. 2.8). Lotka and Price are still widely cited, while Goffman has received less attention throughout the years. The total number of citations is 612 for Lotka's paper of 1926, 73 citations for Goffman's two papers, and 497 for Price's paper from 1965. It should be noted that even though the four seminal papers chosen for the analysis describe models applied specifically to the study and understanding of the science system, none of them were published in Library and Information Science journals. Additionally, only Price is considered a pioneer in the scientific community. His influence results from a series of documents and papers that keep him visible in the scientometric community. Both Derek de Solla Price and Alfred J. Lotka have around 50 papers in the ISI Web of Science, while William Goffman has little more than 25.

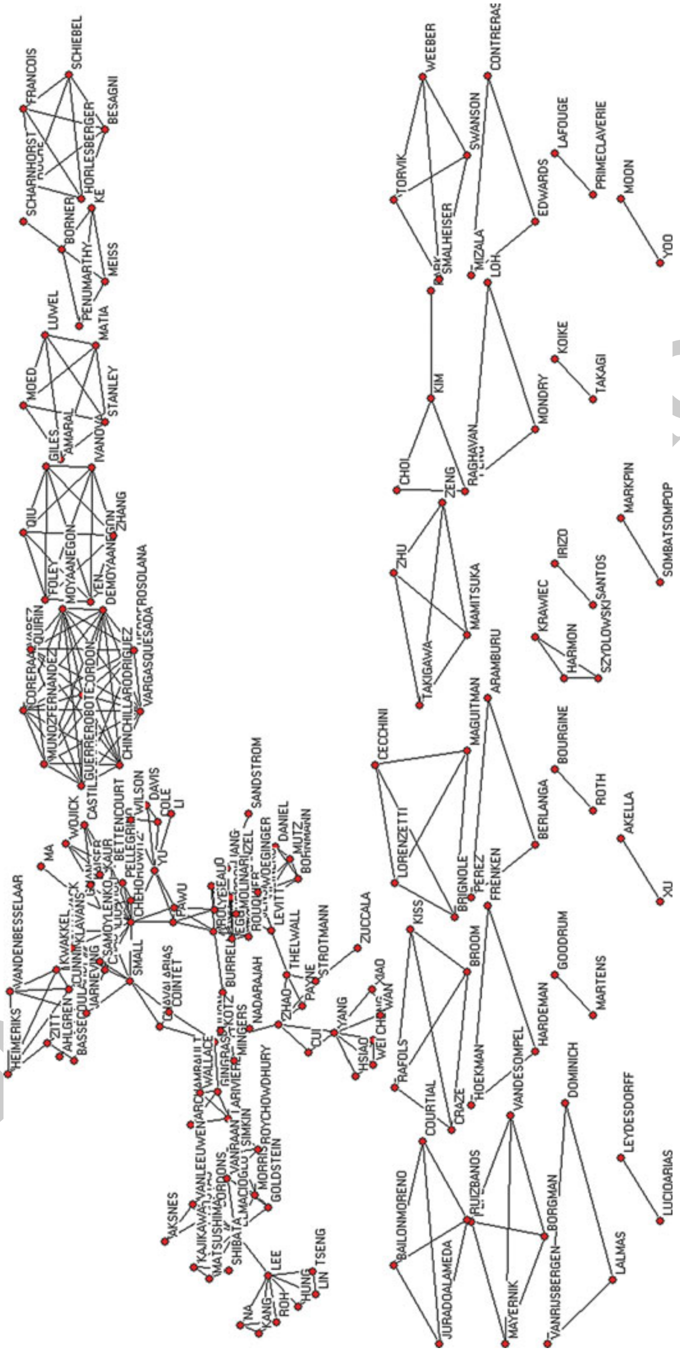
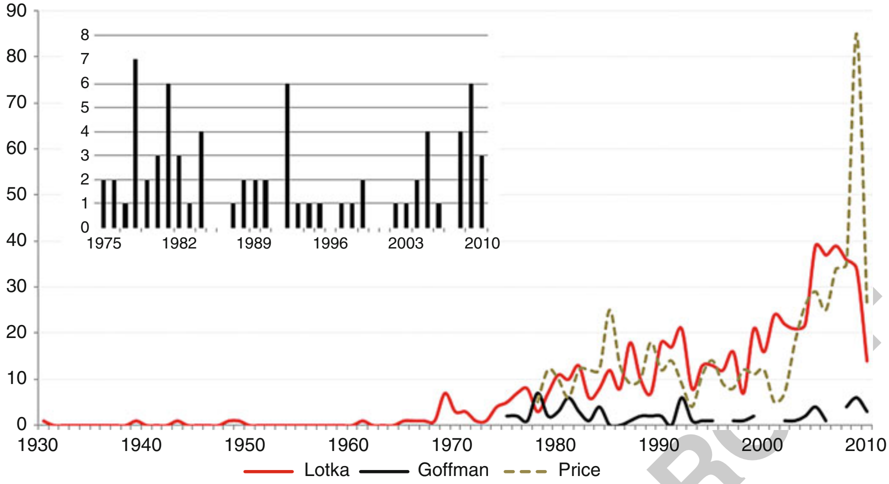


Fig. 2.8 Bibliographic coupling among the 137 documents identified as using mathematical models of science (cosine 0,1)

this figure will be printed in b/w



this figure will be printed in b/w

Fig. 2.9 Growth of papers citing the three historical models – Lotka's law, Goffman's model, and Price's model. The yearly citations of Goffman's two papers are shown within the inlay

Figure 2.9 shows the annual number of citations for three cases. In the case of Lotka, we see that his model is still influential eight decades after its publication, although it took some years for it to become popular in the scientific community. The reception of Price and Lotka (at least of their papers of 1926 and 1965) seems to be similar. Although there is also an underlying growth of the Web of Science, the reception of both papers grows together with the consolidation of scientometrics as a field (Lucio-Arias and Leydesdorff 2009).

For the case of Goffman, there are few documents citing the two selected papers. Therefore, we have displayed the annual citation numbers in an additional figure as an inlay in Fig. 2.9. From this bar chart, we can see that the annual numbers are small, the papers disappear from the radar now and then, and there is a kind of revival of popularity beginning around 2000. With its more robust growth of perception, the Price model also seems to gain popularity after 2000. Actually, both models – Goffman's as well as Price's – have also been discussed together with the emergence of network science and the application of network science to the science system (Börner et al. 2007).

We also display the HistCite^{TM} graphs for all three cases (four papers) for a visual impression. As can be seen from Fig. 2.10, they are quite different in nature. While the graphs are very dense for the case of Lotka's and Price's models, in the case of Goffman's model there are fewer nodes and a more sparsely connected network. We will look into the diffusion pattern in all three cases separately in more detail.

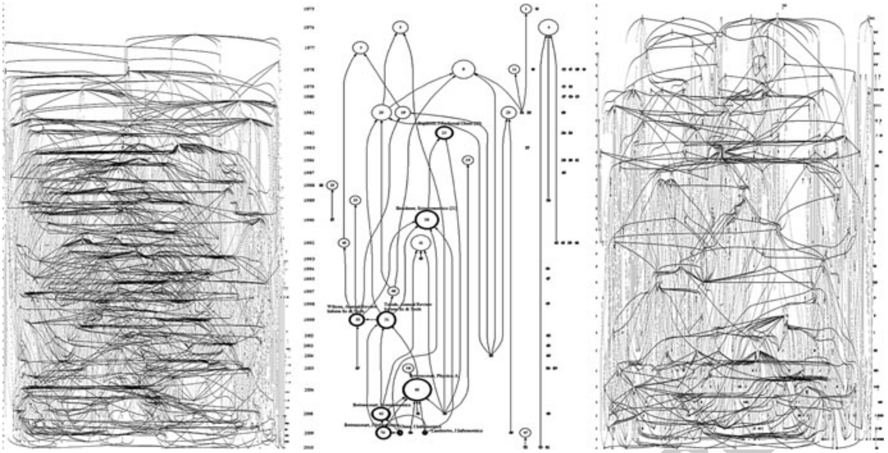


Fig. 2.10 Historiographs for Lotka (*left*), Goffman (*middle*), and Price (*right*) – overview

2.4.2.2 Physics of Evolution: From Biological Species to Productive Actors – A.J. Lotka

Lotka's law reflects a regularity concerning the productivity of scholars (measured by the number of publications). Lotka found that a majority of authors (consisting of a given set of authors) only produce one publication in a given period of time and only very few authors publish larger amounts of articles. If the number of authors with *n* publications is plotted against the aggregated volume of publications, we find an inverted power law with an exponent that is in many cases near 2. Lotka's law is an empirical law with authors as the basic unit of analysis. It is one of the fundamental bibliometric laws that, relatively speaking, can be easily tested against very different bibliometric samples, which explains its overwhelming success. Researchers have discussed how collaboration influences productivity (e.g. Kretschmer and Kretschmer 2007) and how productivity patterns change between different generations of researchers (e.g. Fronczak et al. 2007). But Lotka's law is more than just a statistical regularity. It belongs to a class of mathematical distributions that are characteristic of complex processes not only in social systems, but also in natural systems (Bak 1996). For information processes, even the label of "Lotkaian informetrics" has been used by Egghe in his systematic mathematical analysis of functions used to describe Lotka's law. Lotka's mathematical model is a descriptive one. But it can be used as a litmus test for any predictive model of scientific activity that also entails scientists and publications. For instance, in his agent-based model, through which topics, papers and authors find each other and form scientific fields, Gilbert (1997) calculated Lotka's law to see if his artificial science simulation reveals structures similar to real science.

Details about Lotka's law are given in Chap. 3 of this book. The emphasis here is on its diffusion through the years, the applications of the law, and the characteristics

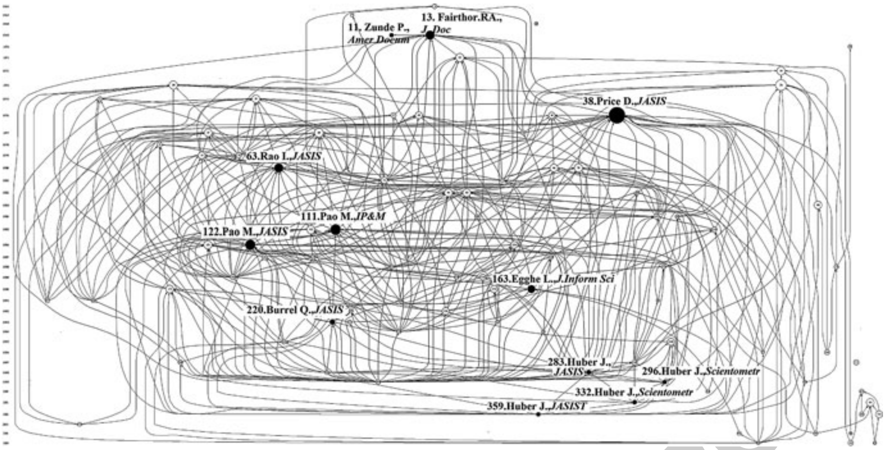


Fig. 2.11 Historiograph of documents citing Lotka's law and main path

of those documents citing it. A total of 612 documents cite "The frequency distribution of scientific productivity," Alfred Lotka's 1926 paper published in the Journal of the Washington Academy of Sciences. The number of publications dealing with the informetric law of the skewed distribution of publications is so large that it is possible to verify Lotka's law using a set of papers devoted to his law of scientific productivity (Yablonsky 1980).

The reconstruction of the diffusion trajectories of Lotka using HistCite$^{\text{TM}}$ (see Fig. 2.10, right) illustrates cohesiveness in the set: authors citing Lotka are also both aware of each other and citing each other. Figure 2.10 also gives an impression of the size and density of the network of papers citing Lotka's paper of 1926 (the graph is not displayed for detailed inspection[5]). Lotka's law is cited in more than 200 different journals, but more than 50% of them correspond to the ISI subject category of Library and Information Science. This way, the graph also reflects the dominance of Scientometrics as part of LIS disciplines inside the set. The graph illustrates how Lotka's law becomes a relevant "knowledge item" that binds papers together in the flows of information and knowledge production and that contributes to a consolidation of scientometrics as a scientific field, for which a high connectivity of networks of citations is one important feature. For a slightly more detailed inspection, we reproduce the historiograph using as a threshold at least five citations from other documents of the set (91 nodes).

In Fig. 2.11, the nodes of the main path or backbone are highlighted and labeled. There is an important volume of documents that either refers to Lotka's formula in a more rhetorical way or discusses mechanisms for and implications of this law in the light of social theories. But most of the documents highlighted by the main path

---

[5] We will provide a on-line version for detailed inspection.

Table 2.3 Main path of documents citing Lotka

First Author	PY	Journal	Title
Zunde, P	1969	JASIST	Indexing consistency and quality
Fairthor, RA	1969	J.DOC.	Progress in documentation – empirical hyperbolic distributions (Bradford–Zipf–Mandelbrot) for bibliometric description and prediction
Price, DJD	1976	JASIST	General theory of bibliometric and other cumulative advantage processes
Rao, IKR	1980	JASIST	Distribution of scientific productivity and social-change
Pao, ML	1985	IP&M	Lotka law – a testing procedure
Pao, ML	1986	JASIS	An empirical-examination of Lotka law
Egghe, L	1990	J. INFORMATION SCIENCE	The duality of informetric systems with applications to the empirical laws
Burrell, Ql	1993	JASIST	Yes, the GIGP really does work – and is workable
Huber, JC	1998	JASIST	Cumulative advantage and success-breeds-success: the value of time pattern analysis
Huber, JC	1999	SCIENTOMETRICS	Inventive productivity and the statistics of exceedances
Huber, JC	2001	SCIENTOMETRICS	Scientific production: a statistical analysis of authors in mathematical logic
Huber, JC	2002	JASIST	A new model that generates Lotka's law

of Fig. 2.11 (*dark circles*) entail mathematical formulations or applications (e.g., for descriptive statistics of research fields, journals, or specific regions or countries). Most of the documents using Lotka's law rely on empirical data at a meso level of aggregation (101–10,000 records). A bibliographic description of the documents belonging to the main path is available in Table 2.3. Most of these papers discuss Lotka's law in the context of specific distribution functions and stochastic processes that lead to them.

2.4.2.3 The Case of Modeling the Spreading of Ideas as a Disease – W. Goffman

Goffman's model describes the spreading out of an idea as analogous to the spreading of a disease. Similar to Lotka's law, which is part of the long history in

Table 2.4 Main path of documents citing Goffman

First Author	PY	Journal	Title
Bujdoso, E	1982	J. RADIOANAL CHEM	Prompt nuclear analysis – growth and trends – a scientometric study
Bruckner, E	1990	SCIENTOMETRICS	The application of evolution models in scientometrics
Wilson, CS	1999	ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY	Informetrics
Tabah, AN	1999	ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY	Literature dynamics: studies on growth, diffusion, and epidemics
Bettencourt, LMA	2006	PHYSICA A	The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models
Bettencourt, LMA	2008	SCIENTOMETRICS	Population modeling of the emergence and development of scientific fields
Lambiotte, R	2009	JOURNAL OF INFORMETRICS	Communities, knowledge creation, and information diffusion
Chen, CM	2009	JOURNAL OF INFORMETRICS	Towards an explanatory and computational theory of scientific discovery
Bettencourt, LMA	2009	JOURNAL OF INFORMETRICS	Scientific discovery and topological transitions in collaboration networks

the study of statistical distributions, the epidemic model Goffman adopted has a long history. In 1927, Kermack and McKendrick published a mathematical model that is still known as the SIR model. This model describes the spreading out of a disease in terms of the relative growth of three subpopulations: the number of susceptible but uninfected individuals (S), the number of infected individuals (I) who carry the disease and can spread it further to the S-group, and the number of recovered individuals (R) who cannot be reinfected again. Obviously, the growth of infected individuals depends on the number of available susceptible individuals and is slowed down by recovering. Goffman applied this idea to science. The number of "infected" researchers represents the researchers working at an idea or in a field. The R-group has lost interest and the S-group forms the reservoir for further growth. Unlike Lotka's law, for which only one key publication can be found, Goffman published work about this model over the course of several years, and also with different co-authors (Harmon 2008). For our analysis, we identified two main publications that still gain sufficient recognition (Table 2.4). AQ3

Goffman's model entails many more variables (three instead of one) and many more parameters than Lotka's law. Although it has been tested empirically

(Wagner-Döbler 1999), the number of "susceptible" researchers is not easy to estimate (Burger and Bujdoso 1985). Nevertheless, one prediction of Goffman's model can easily be measured: the growth of a scientific field. Scientometrics has produced a large amount of growth studies of new scientific fields. Correspondingly, the literature about growth laws in science also makes references to Goffman's model as one possible explanation of such observed growth curves (Tabah 1999). Consequently, Goffman's model has been extended – from the growth of one field (based on the interaction of researchers at three different stages) to the growth of a group of fields (Bruckner et al. 1990). It has also been extended from a group-based model, where the probability of being "infected" with an idea is the same for each subgroup member, to a network-based model, in which the concrete transmission path and the topology of all possible contacts matter (Bettencourt et al. 2009; Lambiotte and Panzarasa 2009).

This history of perception is visible in the main path of the HistCite$^{\text{TM}}$ graph (darker nodes in Fig. 2.12). The 73 citing documents are published in 47 journals illustrating a much more dispersed trajectory of diffusion. Although the Goffman epidemic model is known in the scientometric community, the participation of Library and Information Science journals among the documents citing the seed papers is never as relevant as was the case for Lotka's law.

The main-path analysis also reveals that there is nearly 10-year between the documents in the main path, meaning that once in a decade a paper appears that reminds us of or reviews epidemic models and related approaches. Beginning in 2000, however, the situation changes. Works by Bettencourt et al. (2008, 2009), and later Lambiotte et al. (2009), mark the emergence of the theory of complex-networks in statistical physics (Scharnhorst 2003; Pyka and Scharnhorst 2009). This represents a solid hype, in which new attention from physicists was drawn to the science system.

The science system is a social system for which large (digital) data sets are available. These sets entail a lot of relational information from which different networks can be built and analyzed (Havemann 2009). At the moment, the complex-networks community has shifted its focus from analyzing the structure (as the logical first step of a statistical analysis) to examining the evolution of the network structure (Pastor-Satorras and Vespignani 2004), and further to studying dynamic processes on complex-network topologies. Epidemic modeling has experienced an important revival, and it has been accompanied by a revival of epidemic models of science. The new network science has also influenced the reception of our last case.

2.4.2.4 Network Dynamics from Science and Beyond – Derek de Solla Price

Derek de Solla Price is considered one of the pioneers in the field of Scientometrics. He has written about many different topics, and his work is still highly cited in the scientometric community. In 1965, he published a relatively short paper in the journal Science entitled "Networks of papers." Although this paper contains only a few formulas, it has established a foundation for further study of

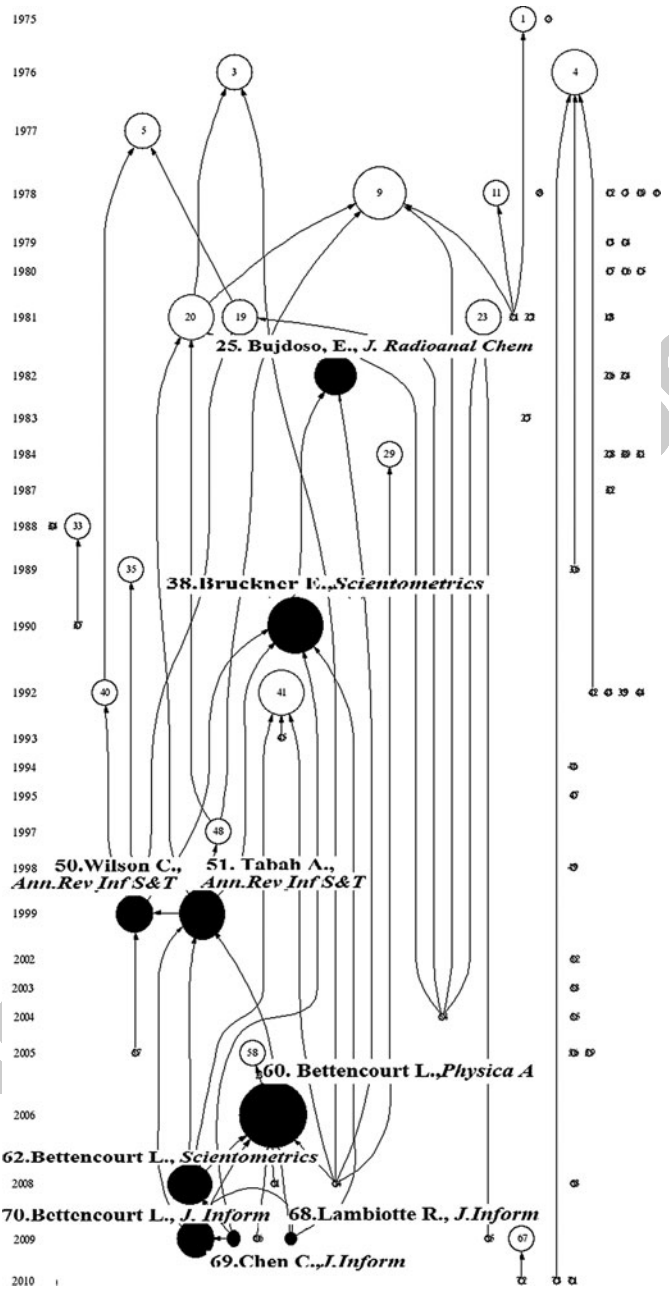


Fig. 2.12 Historiograph of documents citing Goffman's epidemic model

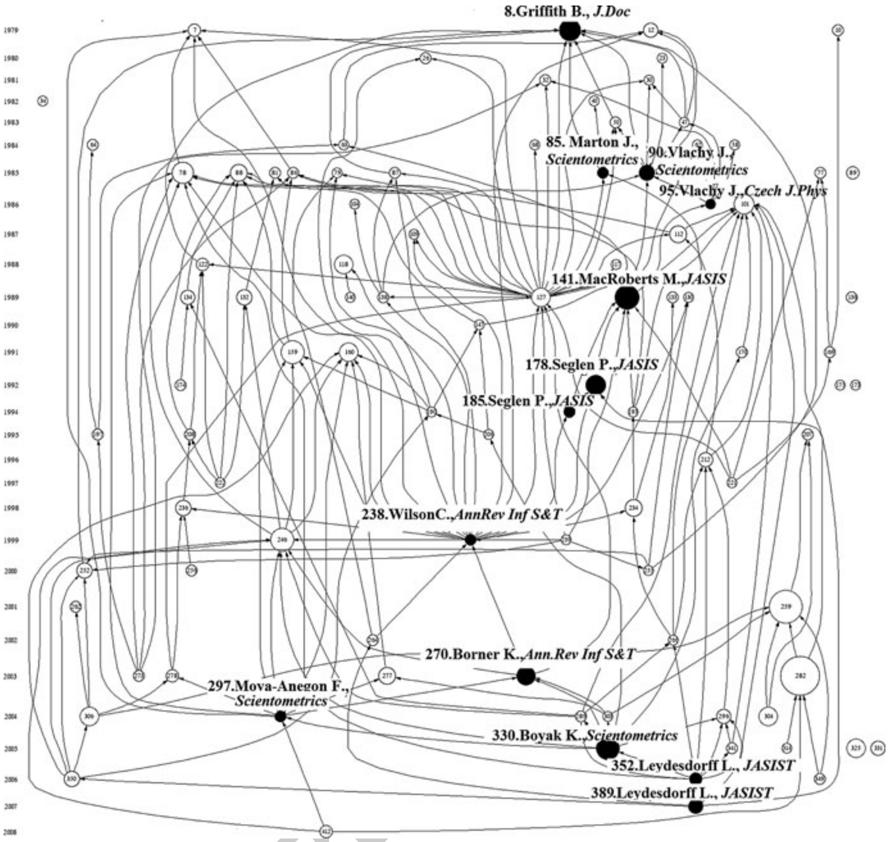


Fig. 2.13 Historiograph of documents citing Price's network model of scientific papers and main path

scientific communication, including mathematical models. Price begins his paper with the observation that citations are skewed in their distribution. He examines the consequences of the (exponential) growth of publications (one of his other major findings) for the future distribution of citations, and he argues that although references and citations form a balance, their distribution over papers differs fundamentally. Citations are not homogeneously distributed over the growing body of literature. Instead, they cluster in time and space (defined as sets of papers). Based on these structures, we can identify research fronts. Citing is the recursive and constitutive process that redefines, reshapes, and re-creates scientific knowledge for each generation of scholars. Price visualizes the evolution of networks of papers. He not only reflects upon fundamental bibliographic questions such as classification, he also points to a number of unknown or unclear characteristics of the self-organized, collective process of references, later addressed by measurements and models.

Table 2.5 Main path of documents citing Price

First Author	PY	Journal	Title
Griffith, BC	1979	J.DOC	Aging of scientific literature – citation analysis
Vlachy, J	1985	SCIENTOMETRICS	Citation histories of scientific publications – the data sources
Marton, J	1985	SCIENTOMETRICS	Obsolescence or immediacy – evidence supporting Price hypothesis
Vlachy, J	1986	CZECH J PHYSICS	Scientometric analyses in physics – where we stand
Macroberts, BR	1989	JASIST	Problems of citation analysis – a critical review
Seglen, PO	1992	JASIST	The skewness of science
Seglen, PO	1994	JASIST	Causal relationship between article citedness and journal impact
Wilson, CS	1999	ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY	Informetrics
Borner, K	2003	ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY	Visualizing knowledge domains
Moya-Anegon, F	2004	SCIENTOMETRICS	A new technique for building maps of large scientific domains based on the cocitation of classes and categories
Boyack, KW	2005	SCIENTOMETRICS	Mapping the backbone of science
Leydesdorff, L	2006	JASIST	Can scientific journals be classified in terms of aggregated journal-journal citation relations using the journal citation reports?
Leydesdorff, L	2007	JASIST	Betweenness centrality as an indicator of the interdisciplinarity of scientific journals

Due to Price's overall relevance to the scientometric community and his rich trajectory of published papers relevant to this field, documents citing Price's network model are mostly published in journals of Library and Information Science. This is similar to the case of Lotka's law. In Price's case, we also present the HistCite^{TM} graph for visual inspection (Fig. 2.13).

The historiograph shown in Fig. 2.13 illustrates a cohesive set of documents similar to the case of Lotka's law. However, the authors citing Price do not possess the same awareness of each other as was for the case for authors using the Lotka model. For this reason, it was possible to lower the threshold used in Lotka's case (citing at least five other documents) to all those documents citing at least three other documents (96 nodes). The network is less dense, justifying a lower threshold. The documents in the main path (*dark labeled nodes* in Fig. 2.13) are detailed in Table 2.5.

A comparison of the backbone of Lotka and Price reinforces the impression that comes with an inspection of all journals in the two data sets. Both authors and both models are part of the knowledge base of scientometrics and are fully embraced by the community. This can still not be said for Goffman, however.

2.5 Concluding Remarks

To a certain extent, the analysis from present to past and from past to present complement each other. We found empirical evidence for the narrative drawn at the beginning of this chapter. In particular, the scattered and partly isolated nature of mathematical approaches could be made visible with the help of citation analysis. We found different schools or threads of mathematical approaches and models in a wide sense in LIS – led by statistical analysis and stochastic processes. But although they all draw on a more widely connected network of mathematical approaches, they do not communicate this among each other. We also found evidence for the still relatively marginal role of dynamic models in the set of current papers in LIS, as well as in the way Goffman (as one of the proponents of dynamic models) is hardly recognized in the LIS community.

Concerning the relation between predictive and descriptive models of science, which is one of the topics addressed by this book (see in particular Chap. 1), our empirical analysis underlines once more that when mathematical models are currently applied to describe the development of science at all, they rather focus on an analysis of the current state in a descriptive way. However, each mathematical model with a dynamic component also has the potential to be applied for prediction. Let us give an example: Lotka's law of productivity is just a mathematical function between variables (number of scientists, number of their publications) that can be empirically tested. This means it is predictive in its essence. However, any stochastic process proposed to explain the establishment of Lotka's law as a quasi-stationary distribution of a dynamic process makes assumptions about micromechanisms of behavior. One possible assumption is that the probability of producing an additional article depends on the number of articles an author has already produced. Such a rule can be implemented in models explicitly designed to test the collective outcome of behavioral rules on the level of individuals (such as Gilbert's model). We can also use such assumptions about micromechanisms and the parameters of Lotka's law to predict the productivity of a certain scientific community. However, only a few attempts have been made to turn mathematical models of science into predictive models for scientific development (see Fronczak et al. 2007). This may have more to do with the actual focus of research agendas than the potential of mathematical models as such.

When talking about “predictive modeling,” what is often expressed is the wish to forecast a new idea or a new field. However, in the history of mathematical models of science, one of the predictive models in posse (Goffman’s epidemic model) has been mainly applied in esse to the history of scientific fields (e.g., (Wagner-Döbler 1999)). There are two reasons for this apparent mismatch. First, innovative ideas and new fields representing “real” breakthroughs cannot be predicted by definition. Otherwise, there would not be structural changes of the whole science system, only minor alterations of existing knowledge. Now, what can be predicted also depends on how we define innovation and new ideas. We might reasonably be able to suggest the directions of incremental scientific progress, but not (as said before) radical innovations. In this respect, predictive models are condemned to fail. Peter Allen used to express it in this way: “The more ‘credible’ predictions are, the more likely they are to NOT happen” (Cited in Ebeling and Scharnhorst (2009)).

Yet, while models might fail to predict actual innovations, they have a great and often overlooked potential to analyze the circumstances under which innovations – new ideas and new fields emerging independently of their essence – will most likely arise. Only some of the modeling attempts in the past figuring in our analysis have discussed this aspect (Bruckner et al. 1990). Understood in this way, the potential of models to predict “innovative sciences” – their collaboration pattern, their selection mechanisms, their institutional frames, and so on – is unlimited, and still unexplored. Within such a frame, both descriptive (or, better, statistical) models and predictive (or, better, dynamical) models can be applied. The first can depict characteristics of successful science in the past and search for similar patterns in the present; the second can formulate hypotheses about mechanisms for successful science, test them empirically in the past, and shape them for the present by means of science policy.

Having pointed to this need of modeling for forecasting conditions of events rather than the events themselves, we immediately have to admit that differentiating and tracing such a use of mathematical models is almost impossible by the analysis of citations only. Again, citation analysis can point us to interesting areas to look at more closely. But for the actual use, application, and interpretation of models, we either have to rely on manual inspection or on other kind of references that relate a model to a certain use. That seems to be even harder to trace semi-automatically than the pure appearance of mathematical models.

What we have done in this analysis is to describe the current state of diffusion of mathematical modeling ideas irrespectively of their actual use. Already, this confronted us with a lot of problems. To trace an adoption pattern as sketched in Fig. 2.1, we would need to be able to automatically extract all documents (across all disciplines) that address the application of the mathematical models to the science system. Moreover, we would also like to see in parallel the bibliometric traces of the mathematical branches feeding these models. However, there is no consistent indexing of documents (outside of knowledge-domain-specific databases) concerning the methods they apply. We also found that there is no term-keyword-subject combination that delivers a specific enough set of documents for mathematical models in science over the whole Web of Science database. This is why we have chosen the combination of tracing known model approaches to science (over all disciplines) with screening a set of established LIS journals for the appearance of mathematical modeling.

Despite this limited-sampling approach and specific-citation perspective, we found evidence both for the relatively isolated existence of mathematical modeling and its implicit commonly shared knowledge base. We also saw the influence of developments in other fields on the implementation of new methods in LIS. The emergence of the so-called new network science (Barabási 2002) and the interest from statistical physics and, in a wider sense, complexity research (all three representing the mother disciplines for dynamic processes) do not remain without resonance in scientometrics. Partly, we observe a diffusion of new researchers; partly, we also observe a taking up of themes and methods by established scientometricians who in some way received their primary academic forming in natural sciences and mathematics.

Our experiments show that developments in scientometrics cannot be understood from an inner-situated perspective only. The use of mathematical dynamic models to describe the sciences is not restricted to LIS journals. Actually, some interesting developments in this area take place at very different locations, such as in journals of computational philosophy (see Chap. 4 of this book), sociology (see Chap. 6 of this book), and physics. But the universal nature of mathematical dynamic approaches – their variety in methods and topics addressed – makes it impossible to set up a string of keywords with which one can easily extract a good sample of mathematical models applied to the science system. The same holds for a past-to-present analysis. Mathematical models applied to science can pop up in all places. We selected three researchers – Lotka, Goffman, and Price – who performed pioneering work relevant to scientometrics, who have been interested in dynamic processes, and who have developed mathematical models and/or ideas that have been central for modeling. There might be many other researchers who have done interesting modeling experiments and might only be rediscovered by chance. But even for our three "landmark" scholars, it is not easy for us to pick one publication from their oeuvre that fully represents their "science model" and nothing else. The work of an individual scholar is like a journey through a landscape of science. Partly discovering the existing landscape for her/himself and partly creating this landscape, the scholar leaves marks and traces and is marked and imprinted by their journey. One might argue that there is a certain arbitrariness in the selection of our cases and the seed nodes for the historiographic methods. Indeed, we are aware of this. We do not claim comprehensiveness; instead, we aim for an insightful illustration of the complexity of knowledge and model transfer in science. Our practical problems in the selection of samples also reflect a more fundamental problem.

The diffusion of ideas and methods across the sciences is a combination of the progress of knowledge inside specialties and a diffusion of knowledge between specialties in which knowledge is not just transmitted but also altered. The evolution of knowledge entails processes of specification as well as generalization. Correspondingly, in the cognitive and social space, specialties and invisible colleges emerge and disappear, merge and split up, take form, stabilize, transform, and

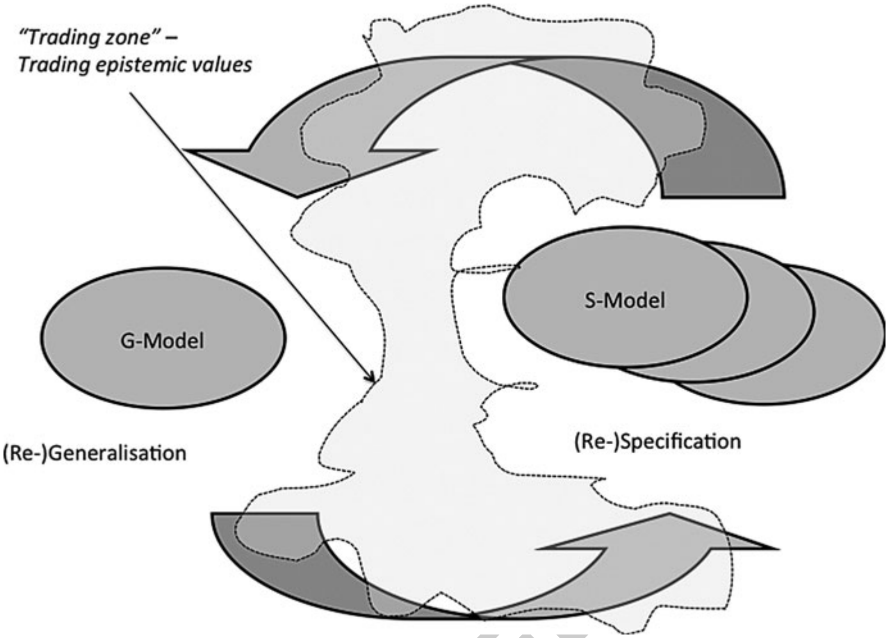


Fig. 2.14 Models travelling between generic and specific levels

pass different stages of life cycles, all based on a constant flow of creation and recombination of elementary units of knowledge. This eternal flow also influences the travel of mathematical models. Approaches to mathematical modeling can emerge on a generic mathematical level or inside of a specialty or knowledge domain. Independently from where they appear first, they are embedded in a cycle of (re)generalization and (re)specification (Fig. 2.14). One of these special fields can be scientometrics. Mathematical models can be developed specifically for science. However, they will always share a generic structural element with other models and contribute to this pool. On the other side, from the general pool of models they can expect entries of new model ideas along all possible lines of mathematical modeling. Mathematical models and approaches to science can be the result of applying different mathematical approaches that have been used in other disciplines. For example, some models using entropy statistics stem from the Mathematical Theory of Communications, which originally addressed an engineering problem but which has been applied in more social sciences like economics.

This feature of the model-building process – the cycle between generalization and specification – makes it very complicated to trace a model transfer bibliometrically. It also makes it hard to produce an overview of possible dynamic models of science, which in principle encompass all dynamic modeling approaches.

Therefore, we applied a practical approach by concentrating on LIS journals for the analysis of the present situation and by depicting a few "classics" from the past.

The combination of both approaches provides bibliometric evidence for less cited mathematical approaches that have been fading away, for models that have been only recently (re)discovered, and for a shared underlying cognitive reference space that is not always visible in direct citations. Our study also illustrates the process of spreading new ideas and demonstrates how these can eventually converge. It can be expected that such a historiographic study can be used as a departure point for an evaluation of certain mathematical models. What are the characteristics of the most successful models? Do they tend to be more universal or domain specific? Are they multi-leveled? We can also imagine applying some of the characteristics of models discussed in the Introduction Chapter in a future analysis. For instance, one could ask about the quantitative or qualitative nature of the models applied, the type of behavior in science targeted, and the representation used for results.

Last but not least, one remark. In our historic narrative at the beginning of this chapter, we argued that eventually there need to be researchers who are intrigued and curious enough to test mathematical models. However, while researchers as the source of ideas remain utterly important, mathematical modeling will still remain ephemeral if it is to be an activity driven by curiosity and not by demand. The creativity of the human imagination is triggered by curiosity as well as by a societal demand for a certain type of knowledge, method, and models. There is no sustainable modeling without a thorough theoretical foundation, and, in this respect, models should be mainly guided by theory.

One could argue that, compared to other fields and disciplines, scientometrics is a relatively young field and has therefore not yet penetrated or been open to complex models very much. But dynamic modeling of the science system will not emerge if there is not a need to apply relatively complex, computational-intensive models that also require diverse collaborations. The pertinent growth of the science system, the scarcity of resources (human and material), and the increasing complexity that requires other mechanisms of control might all be decisive in triggering a collective action for Modeling Science Dynamics.

Appendix 1: Papers Using Mathematical Approaches to Understand the Science System (Fig. 2.1)

AQ4

Table 2.6 Statistics of the search: Present to past

Node	Bibliographic metadata	Times cited
1	Torvik VI, 2005, J AM SOC INF SCI TECHNOL, V56, P140	18
2	Xekalaki E, 2005, SCIENTOMETRICS, V62, P293	1
3	Santos JB, 2005, SCIENTOMETRICS, V62, P329	2
4	Simkin MV, 2005, SCIENTOMETRICS, V62, P367	8
5	Bailon-Moreno R, 2005, SCIENTOMETRICS, V63, P231	1

(continued)

Table 2.6 (continued)

Node	Bibliographic metadata	Times cited
6	Egghe L, 2005, J AM SOC INF SCI TECHNOL, V56, P664	2
7	Egghe L, 2005, J AM SOC INF SCI TECHNOL, V56, P669	8
8	Sombatsompop N, 2005, J AM SOC INF SCI TECHNOL, V56, P676	15
9	Matia K, 2005, J AM SOC INF SCI TECHNOL, V56, P893	10
10	Egghe L, 2005, J AM SOC INF SCI TECHNOL, V56, P935	13
11	Efron M, 2005, J AM SOC INF SCI TECHNOL, V56, P969	4
12	Liang LM, 2005, J AM SOC INF SCI TECHNOL, V56, P1045	10
13	Burrell QL, 2005, SCIENTOMETRICS, V64, P247	5
14	Morris SA, 2005, J AM SOC INF SCI TECHNOL, V56, P1250	5
15	Burrell QL, 2005, INFORM PROCESS MANAGE, V41, P1317	8
16	Egghe L, 2005, INFORM PROCESS MANAGE, V41, P1330	2
17	Shan S, 2005, INFORM PROCESS MANAGE, V41, P1369	2
18	Lafouge T, 2005, INFORM PROCESS MANAGE, V41, P1387	3
19	Payne N, 2005, INFORM PROCESS MANAGE, V41, P1495	1
20	Coccia M, 2005, SCIENTOMETRICS, V65, P307	0
21	Burrell QL, 2005, SCIENTOMETRICS, V65, P381	11
22	Dominich S, 2006, INFORM PROCESS MANAGE, V42, P1	0
23	Zuccala A, 2006, J AM SOC INF SCI TECHNOL, V57, P152	14
24	Aksnes DW, 2006, J AM SOC INF SCI TECHNOL, V57, P169	13
25	Klavans R, 2006, J AM SOC INF SCI TECHNOL, V57, P251	33
26	Ackermann E, 2006, SCIENTOMETRICS, V66, P451	0
27	Martens BVD, 2006, J AM SOC INF SCI TECHNOL, V57, P330	1
28	Chen CM, 2006, J AM SOC INF SCI TECHNOL, V57, P359	67
29	van Raan AFJ, 2006, J AM SOC INF SCI TECHNOL, V57, P408	21
30	Choi J, 2006, INFORM PROCESS MANAGE, V42, P331	5
31	Wei CP, 2006, INFORM PROCESS MANAGE, V42, P350	12
32	Izsak J, 2006, SCIENTOMETRICS, V67, P107	2
33	Yoo SH, 2006, SCIENTOMETRICS, V69, P57	0
34	Egghe L, 2006, SCIENTOMETRICS, V69, P121	67
35	Glanzel W, 2006, SCIENTOMETRICS, V67, P315	73
36	Burrell QL, 2006, SCIENTOMETRICS, V67, P323	0
37	Rousseau R, 2006, J AM SOC INF SCI TECHNOL, V57, P1404	2
38	Burrell QL, 2006, J AM SOC INF SCI TECHNOL, V57, P1406	4
39	Samoylenko I, 2006, J AM SOC INF SCI TECHNOL, V57, P1461	7
40	Peng D, 2006, SCIENTOMETRICS, V69, P271	0
41	Roth C, 2006, SCIENTOMETRICS, V69, P429	5
42	Mingers J, 2006, INFORM PROCESS MANAGE, V42, P1451	6
43	Zitt M, 2006, INFORM PROCESS MANAGE, V42, P1513	17
44	Su Y, 2006, J AM SOC INF SCI TECHNOL, V57, P1977	0
45	Van Den Besselaar P, 2006, SCIENTOMETRICS, V68, P377	10
46	Borner K, 2006, SCIENTOMETRICS, V68, P415	8
47	Klavans R, 2006, SCIENTOMETRICS, V68, P475	15
48	Small H, 2006, SCIENTOMETRICS, V68, P595	23
49	Contreras C, 2006, SCIENTOMETRICS, V69, P689	3
50	Kretschmer H, 2007, J INFORMETR, V1, P308	1
51	Jarneving B, 2007, J INFORMETR, V1, P338	0
52	Burrell QL, 2007, J INFORMETR, V1, P16	22

(continued)

Table 2.6 (continued)

Node	Bibliographic metadata	Times cited
53	Leydesdorff L, 2007, J AM SOC INF SCI TECHNOL, V58, P25	14
54	McDonald JD, 2007, J AM SOC INF SCI TECHNOL, V58, P39	4
55	Koike A, 2007, J AM SOC INF SCI TECHNOL, V58, P51	2
56	Egghe L, 2007, J AM SOC INF SCI TECHNOL, V58, P452	29
57	Na SH, 2007, INFORM PROCESS MANAGE, V43, P302	1
58	Lucio-Arias D, 2007, SCIENTOMETRICS, V70, P603	3
59	Egghe L, 2007, J INFORMETR, V1, P115	5
60	Shibata N, 2007, J AM SOC INF SCI TECHNOL, V58, P872	2
61	Zitt M, 2007, INFORM PROCESS MANAGE, V43, P834	0
62	Zhao DZ, 2007, J AM SOC INF SCI TECHNOL, V58, P1285	2
63	Rousseau R, 2007, J AM SOC INF SCI TECHNOL, V58, P1551	3
64	Nadarajah S, 2007, SCIENTOMETRICS, V72, P291	1
65	Simkin MV, 2007, J AM SOC INF SCI TECHNOL, V58, P1661	6
66	Morris SA, 2007, J AM SOC INF SCI TECHNOL, V58, P1764	4
67	Bornmann L, 2007, SCIENTOMETRICS, V73, P139	1
68	de Moya-Anegon F, 2007, J AM SOC INF SCI TECHNOL, V58, P2167	9
69	Lariviere V, 2008, J AM SOC INF SCI TECHNOL, V59, P288	12
70	van Raan AFJ, 2008, J AM SOC INF SCI TECHNOL, V59, P565	7
71	Bornmann L, 2008, J AM SOC INF SCI TECHNOL, V59, P830	38
72	Chavalarias D, 2008, SCIENTOMETRICS, V75, P37	3
73	Li XY, 2008, INFORM PROCESS MANAGE, V44, P991	0
74	Wan XJ, 2008, INFORM PROCESS MANAGE, V44, P1032	3
75	Molinari A, 2008, SCIENTOMETRICS, V75, P339	7
76	Bettencourt LMA, 2008, SCIENTOMETRICS, V75, P495	7
77	Kim H, 2008, SCIENTOMETRICS, V75, P535	2
78	Harmon G, 2008, INFORM PROCESS MANAGE, V44, P1634	0
79	Bornmann L, 2008, J INFORMETR, V2, P217	2
80	Yu HR, 2008, J INFORMETR, V2, P240	0
81	Egghe L, 2008, SCIENTOMETRICS, V76, P117	4
82	Egghe L, 2008, J AM SOC INF SCI TECHNOL, V59, P1608	13
83	van Raan AFJ, 2008, J AM SOC INF SCI TECHNOL, V59, P1631	3
84	Egghe L, 2008, J AM SOC INF SCI TECHNOL, V59, P1688	2
85	Ahlgren P, 2008, SCIENTOMETRICS, V76, P273	3
86	Burrell QL, 2008, INFORM PROCESS MANAGE, V44, P1794	1
87	Bornmann L, 2008, J INFORMETR, V2, P280	0
88	Ye FY, 2008, J INFORMETR, V2, P288	2
89	Quirin A, 2008, J AM SOC INF SCI TECHNOL, V59, P1912	2
90	Lucio-Arias D, 2008, J AM SOC INF SCI TECHNOL, V59, P1948	3
91	Levitt JM, 2008, J AM SOC INF SCI TECHNOL, V59, P1973	3
92	Cecchini RL, 2008, INFORM PROCESS MANAGE, V44, P1863	1
93	Egghe L, 2008, J AM SOC INF SCI TECHNOL, V59, P2133	0
94	Davis PM, 2008, J AM SOC INF SCI TECHNOL, V59, P2186	6
95	Egghe L, 2008, SCIENTOMETRICS, V77, P377	3
96	Szydlowski M, 2009, SCIENTOMETRICS, V78, P99	0
97	Wallace ML, 2009, J AM SOC INF SCI TECHNOL, V60, P240	1
98	Jensen P, 2009, SCIENTOMETRICS, V78, P467	0

(continued)

Table 2.6 (continued)

Node	Bibliographic metadata	Times cited
99	Costas R, 2009, J AM SOC INF SCI TECHNOL, V60, P740	3
100	Perez JM, 2009, INFORM PROCESS MANAGE, V45, P356	0
101	Sandstrom U, 2009, SCIENTOMETRICS, V79, P341	1
102	Borner K, 2009, J INFORMETR, V3, P161	0
103	Chen CM, 2009, J INFORMETR, V3, P191	2
104	Bettencourt LMA, 2009, J INFORMETR, V3, P210	4
105	Frenken K, 2009, J INFORMETR, V3, P222	2
106	Skupin A, 2009, J INFORMETR, V3, P233	2
107	Lucio-Arias D, 2009, J INFORMETR, V3, P261	6
108	Zhao YY, 2009, SCIENTOMETRICS, V80, P91	0
109	Elmacioglu E, 2009, SCIENTOMETRICS, V80, P195	0
110	Zhu SF, 2009, INFORM PROCESS MANAGE, V45, P555	1
111	Deineko VG, 2009, SCIENTOMETRICS, V80, P819	0
112	Egghe L, 2009, J INFORMETR, V3, P290	2
113	Wallace ML, 2009, J INFORMETR, V3, P296	2
114	Yu LP, 2009, J INFORMETR, V3, P304	0
115	Kwakkel JH, 2009, J AM SOC INF SCI TECHNOL, V60, P2064	0
116	He ZL, 2009, J AM SOC INF SCI TECHNOL, V60, P2151	0
117	Tseng YH, 2009, SCIENTOMETRICS, V81, P73	1
118	Egghe L, 2009, J AM SOC INF SCI TECHNOL, V60, P2362	1
119	Bornmann L, 2009, SCIENTOMETRICS, V81, P407	1
120	Ye FY, 2009, SCIENTOMETRICS, V81, P493	0
121	Egghe L, 2009, SCIENTOMETRICS, V81, P567	0
122	Lucio-Arias D, 2009, J AM SOC INF SCI TECHNOL, V60, P2488	0
123	Luk R, 2009, J AM SOC INF SCI TECHNOL, V60, P2587	0
124	Guan JC, 2009, SCIENTOMETRICS, V81, P683	0
125	Kiss IZ, 2010, J INFORMETR, V4, P74	0
126	Bornmann L, 2010, J INFORMETR, V4, P83	1
127	Guan JC, 2010, SCIENTOMETRICS, V82, P165	0
128	Egghe L, 2010, SCIENTOMETRICS, V82, P243	0
129	Yu G, 2010, SCIENTOMETRICS, V82, P249	0
130	Xu ZB, 2010, INFORM PROCESS MANAGE, V46, P143	0
131	Pepe A, 2010, J AM SOC INF SCI TECHNOL, V61, P567	0
132	Liang LM, 2010, J INFORMETR, V4, P201	0
133	Minguillo D, 2010, J AM SOC INF SCI TECHNOL, V61, P772	0
134	Zhang HZ, 2010, J AM SOC INF SCI TECHNOL, V61, P964	0
135	Egghe L, 2010, SCIENTOMETRICS, V83, P455	0
136	Wray KB, 2010, SCIENTOMETRICS, V83, P471	0
137	Schiebel E, 2010, SCIENTOMETRICS, V83, P765	0

Appendix 2: Cognitive Historiography of Papers Using Mathematical Approaches to Understand the Science System (Fig. 2.2)

Table 2.7 Statistics of the search: Present to past

Node	Bibliographic metadata	Times cited
1	Lotka AJ, 1926, J WASHINGTON ACADEMY, V16, P317	317
2	PRICE DJD, 1965, SCIENCE, V149, P510	664
3	PRICE DJD, 1976, J AMER SOC INFORM SCI, V27, P292	332
4	Barabási AL, 1999, SCIENCE, V286, P509	4818
5	Albert R, 2002, REV MOD PHYS, V74, P47	4030
6	Torvik VI, 2005, J AM SOC INF SCI TECHNOL, V56, P140	18
7	Xekalaki E, 2005, SCIENTOMETRICS, V62, P293	1
8	Santos JB, 2005, SCIENTOMETRICS, V62, P329	2
9	Simkin MV, 2005, SCIENTOMETRICS, V62, P367	8
10	Bailon-Moreno R, 2005, SCIENTOMETRICS, V63, P231	1
11	Egghe L, 2005, J AM SOC INF SCI TECHNOL, V56, P664	2
12	Egghe L, 2005, J AM SOC INF SCI TECHNOL, V56, P669	8
13	Sombatsompop N, 2005, J AM SOC INF SCI TECHNOL, V56, P676	15
14	Matia K, 2005, J AM SOC INF SCI TECHNOL, V56, P893	10
15	Egghe L, 2005, J AM SOC INF SCI TECHNOL, V56, P935	13
16	Efron M, 2005, J AM SOC INF SCI TECHNOL, V56, P969	4
17	Liang LM, 2005, J AM SOC INF SCI TECHNOL, V56, P1045	10
18	Burrell QL, 2005, SCIENTOMETRICS, V64, P247	5
19	Morris SA, 2005, J AM SOC INF SCI TECHNOL, V56, P1250	5
20	Hirsch JE, 2005, PROC NAT ACAD SCI USA, V102, P16569	549
21	Burrell QL, 2005, INFORM PROCESS MANAGE, V41, P1317	8
22	Egghe L, 2005, INFORM PROCESS MANAGE, V41, P1330	2
23	Shan S, 2005, INFORM PROCESS MANAGE, V41, P1369	2
24	Lafouge T, 2005, INFORM PROCESS MANAGE, V41, P1387	3
25	Payne N, 2005, INFORM PROCESS MANAGE, V41, P1495	1
26	Coccia M, 2005, SCIENTOMETRICS, V65, P307	0
27	Burrell QL, 2005, SCIENTOMETRICS, V65, P381	11
28	Bornmann L, 2005, SCIENTOMETRICS, V65, P391	86
29	Dominich S, 2006, INFORM PROCESS MANAGE, V42, P1	0
30	Zuccala A, 2006, J AM SOC INF SCI TECHNOL, V57, P152	14
31	Aksnes DW, 2006, J AM SOC INF SCI TECHNOL, V57, P169	13
32	Klavans R, 2006, J AM SOC INF SCI TECHNOL, V57, P251	33
33	Ackermann E, 2006, SCIENTOMETRICS, V66, P451	0
34	Martens BVD, 2006, J AM SOC INF SCI TECHNOL, V57, P330	1
35	Chen CM, 2006, J AM SOC INF SCI TECHNOL, V57, P359	67
36	van Raan AFJ, 2006, J AM SOC INF SCI TECHNOL, V57, P408	21
37	Choi J, 2006, INFORM PROCESS MANAGE, V42, P331	5
38	Wei CP, 2006, INFORM PROCESS MANAGE, V42, P350	12
39	Izsak J, 2006, SCIENTOMETRICS, V67, P107	2
40	Yoo SH, 2006, SCIENTOMETRICS, V69, P57	0

(continued)

Table 2.7 (continued)

Node	Bibliographic metadata	Times cited
41	Egghe L, 2006, SCIENTOMETRICS, V69, P121	67
42	Glanzel W, 2006, SCIENTOMETRICS, V67, P315	73
43	Burrell QL, 2006, SCIENTOMETRICS, V67, P323	0
44	Rousseau R, 2006, J AM SOC INF SCI TECHNOL, V57, P1404	2
45	Burrell QL, 2006, J AM SOC INF SCI TECHNOL, V57, P1406	4
46	Samoylenko I, 2006, J AM SOC INF SCI TECHNOL, V57, P1461	7
47	Peng D, 2006, SCIENTOMETRICS, V69, P271	0
48	Roth C, 2006, SCIENTOMETRICS, V69, P429	5
49	Mingers J, 2006, INFORM PROCESS MANAGE, V42, P1451	6
50	Zitt M, 2006, INFORM PROCESS MANAGE, V42, P1513	17
51	Su Y, 2006, J AM SOC INF SCI TECHNOL, V57, P1977	0
52	Van Den Besselaar P, 2006, SCIENTOMETRICS, V68, P377	10
53	Borner K, 2006, SCIENTOMETRICS, V68, P415	8
54	Klavans R, 2006, SCIENTOMETRICS, V68, P475	15
55	Small H, 2006, SCIENTOMETRICS, V68, P595	23
56	Contreras C, 2006, SCIENTOMETRICS, V69, P689	3
57	Kretschmer H, 2007, J INFORMETR, V1, P308	1
58	Jarneving B, 2007, J INFORMETR, V1, P338	0
59	Burrell QL, 2007, J INFORMETR, V1, P16	22
60	Leydesdorff L, 2007, J AM SOC INF SCI TECHNOL, V58, P25	14
61	McDonald JD, 2007, J AM SOC INF SCI TECHNOL, V58, P39	4
62	Koike A, 2007, J AM SOC INF SCI TECHNOL, V58, P51	2
63	Egghe L, 2007, J AM SOC INF SCI TECHNOL, V58, P452	29
64	Na SH, 2007, INFORM PROCESS MANAGE, V43, P302	1
65	Lucio-Arias D, 2007, SCIENTOMETRICS, V70, P603	3
66	Egghe L, 2007, J INFORMETR, V1, P115	5
67	Shibata N, 2007, J AM SOC INF SCI TECHNOL, V58, P872	2
68	Zitt M, 2007, INFORM PROCESS MANAGE, V43, P834	0
69	Zhao DZ, 2007, J AM SOC INF SCI TECHNOL, V58, P1285	2
70	Rousseau R, 2007, J AM SOC INF SCI TECHNOL, V58, P1551	3
71	Nadarajah S, 2007, SCIENTOMETRICS, V72, P291	1
72	Simkin MV, 2007, J AM SOC INF SCI TECHNOL, V58, P1661	6
73	Morris SA, 2007, J AM SOC INF SCI TECHNOL, V58, P1764	4
74	Bornmann L, 2007, SCIENTOMETRICS, V73, P139	1
75	de Moya-Anegon F, 2007, J AM SOC INF SCI TECHNOL, V58, P2167	9
76	Lariviere V, 2008, J AM SOC INF SCI TECHNOL, V59, P288	12
77	van Raan AFJ, 2008, J AM SOC INF SCI TECHNOL, V59, P565	7
78	Bornmann L, 2008, J AM SOC INF SCI TECHNOL, V59, P830	38
79	Chavalarias D, 2008, SCIENTOMETRICS, V75, P37	3
80	Li XY, 2008, INFORM PROCESS MANAGE, V44, P991	0
81	Wan XJ, 2008, INFORM PROCESS MANAGE, V44, P1032	3
82	Molinari A, 2008, SCIENTOMETRICS, V75, P339	7
83	Bettencourt LMA, 2008, SCIENTOMETRICS, V75, P495	7
84	Kim H, 2008, SCIENTOMETRICS, V75, P535	2
85	Harmon G, 2008, INFORM PROCESS MANAGE, V44, P1634	0
86	Bornmann L, 2008, J INFORMETR, V2, P217	2

(continued)

Table 2.7 (continued)

Node	Bibliographic metadata	Times cited
87	Yu HR, 2008, J INFORMETR, V2, P240	0
88	Egghe L, 2008, SCIENTOMETRICS, V76, P117	4
89	Egghe L, 2008, J AM SOC INF SCI TECHNOL, V59, P1608	13
90	van Raan AFJ, 2008, J AM SOC INF SCI TECHNOL, V59, P1631	3
91	Egghe L, 2008, J AM SOC INF SCI TECHNOL, V59, P1688	2
92	Ahlgren P, 2008, SCIENTOMETRICS, V76, P273	3
93	Burrell QL, 2008, INFORM PROCESS MANAGE, V44, P1794	1
94	Bornmann L, 2008, J INFORMETR, V2, P280	0
95	Ye FY, 2008, J INFORMETR, V2, P288	2
96	Quirin A, 2008, J AM SOC INF SCI TECHNOL, V59, P1912	2
97	Lucio-Arias D, 2008, J AM SOC INF SCI TECHNOL, V59, P1948	3
98	Levitt JM, 2008, J AM SOC INF SCI TECHNOL, V59, P1973	3
99	Cecchini RL, 2008, INFORM PROCESS MANAGE, V44, P1863	1
100	Egghe L, 2008, J AM SOC INF SCI TECHNOL, V59, P2133	0
101	Davis PM, 2008, J AM SOC INF SCI TECHNOL, V59, P2186	6
102	Egghe L, 2008, SCIENTOMETRICS, V77, P377	3
103	Szydlowski M, 2009, SCIENTOMETRICS, V78, P99	0
104	Wallace ML, 2009, J AM SOC INF SCI TECHNOL, V60, P240	1
105	Jensen P, 2009, SCIENTOMETRICS, V78, P467	0
106	Costas R, 2009, J AM SOC INF SCI TECHNOL, V60, P740	3
107	Perez JM, 2009, INFORM PROCESS MANAGE, V45, P356	0
108	Sandstrom U, 2009, SCIENTOMETRICS, V79, P341	1
109	Borner K, 2009, J INFORMETR, V3, P161	0
110	Chen CM, 2009, J INFORMETR, V3, P191	2
111	Bettencourt LMA, 2009, J INFORMETR, V3, P210	4
112	Frenken K, 2009, J INFORMETR, V3, P222	2
113	Skupin A, 2009, J INFORMETR, V3, P233	2
114	Lucio-Arias D, 2009, J INFORMETR, V3, P261	6
115	Zhao YY, 2009, SCIENTOMETRICS, V80, P91	0
116	Elmacioglu E, 2009, SCIENTOMETRICS, V80, P195	0
117	Zhu SF, 2009, INFORM PROCESS MANAGE, V45, P555	1
118	Deineko VG, 2009, SCIENTOMETRICS, V80, P819	0
119	Egghe L, 2009, J INFORMETR, V3, P290	2
120	Wallace ML, 2009, J INFORMETR, V3, P296	2
121	Yu LP, 2009, J INFORMETR, V3, P304	0
122	Kwakkel JH, 2009, J AM SOC INF SCI TECHNOL, V60, P2064	0
123	He ZL, 2009, J AM SOC INF SCI TECHNOL, V60, P2151	0
124	Tseng YH, 2009, SCIENTOMETRICS, V81, P73	1
125	Egghe L, 2009, J AM SOC INF SCI TECHNOL, V60, P2362	1
126	Bornmann L, 2009, SCIENTOMETRICS, V81, P407	1
127	Ye FY, 2009, SCIENTOMETRICS, V81, P493	0
128	Egghe L, 2009, SCIENTOMETRICS, V81, P567	0
129	Lucio-Arias D, 2009, J AM SOC INF SCI TECHNOL, V60, P2488	0
130	Luk R, 2009, J AM SOC INF SCI TECHNOL, V60, P2587	0
131	Guan JC, 2009, SCIENTOMETRICS, V81, P683	0
132	Kiss IZ, 2010, J INFORMETR, V4, P74	0

(continued)

Table 2.7 (continued)

Node	Bibliographic metadata	Times cited
133	Bornmann L, 2010, J INFORMETR, V4, P83	1
134	Guan JC, 2010, SCIENTOMETRICS, V82, P165	0
135	Egghe L, 2010, SCIENTOMETRICS, V82, P243	0
136	Yu G, 2010, SCIENTOMETRICS, V82, P249	0
137	Xu ZB, 2010, INFORM PROCESS MANAGE, V46, P143	0
138	Pepe A, 2010, J AM SOC INF SCI TECHNOL, V61, P567	0
139	Liang LM, 2010, J INFORMETR, V4, P201	0
140	Minguillo D, 2010, J AM SOC INF SCI TECHNOL, V61, P772	0
141	Zhang HZ, 2010, J AM SOC INF SCI TECHNOL, V61, P964	0
142	Egghe L, 2010, SCIENTOMETRICS, V83, P455	0
143	Wray KB, 2010, SCIENTOMETRICS, V83, P471	0
144	Schiebel E, 2010, SCIENTOMETRICS, V83, P765	0

Appendix 3: Papers from Threads in Figs. 2.3–2.6

Table 2.8 Documents in Fig. 2.3

Node	Author	Year	Journal	Title
4	Simkin, MV	2005	SCIENTOMETRICS	Stochastic modeling of citation slips
10	Egghe, L	2005	JASIST	Zipfian and Lotkaian continuous concentration theory
12	Liang, LM	2005	JASIST	R-sequences: Relative indicators for the rhythm of science
15	Burrell, QL	2005	INFORMATION PROCESSING & MANAGEMENT	Symmetry and other transformation features of Lorenz/Leimkuhler representations of informetric data
21	Burrell, QL	2005	SCIENTOMETRICS	Are "sleeping beauties" to be expected?
29	van Raan, AFJ	2006	JASIST	Statistical properties of Bibliometric indicators: Research group indicator distributions and correlations
38	Burrell, QL	2006	JASIST	On Egghe's version of continuous concentration theory
42	Mingers, J	2006	INFORMATION PROCESSING & MANAGEMENT	Modeling citation behavior in Management Science journals

(continued)

Table 2.8 (continued)

Node	Author	Year	Journal	Title
63	Rousseau, R	2007	JASIST	On Egghe's construction of Lorenz curves
64	Nadarajah, S	2007	SCIENTOMETRICS	Models for citation behavior
65	Simkin, MV	2007	JASIST	A mathematical theory of citing
69	Lariviere, V	2008	JASIST	Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004)
70	van Raan, AFJ	2008	JASIST	Self-citation as an impact-reinforcing mechanism in the science system
83	van Raan, AFJ	2008	JASIST	Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of research groups

Table 2.9 Documents in Fig. 2.4

Node	Author	Year	Journal	Title
25	Klavans, R	2006	JASIST	Identifying a better measure of relatedness for mapping science
28	Chen, CM	2006	JASIST	CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature
45	Van Den Besselaar, P	2006	SCIENTOMETRICS	Mapping research topics using word-reference co-occurrences: A method and an exploratory case study
46	Borner, K	2006	SCIENTOMETRICS	Mapping the diffusion of scholarly knowledge among major US research institutions
47	Klavans, R	2006	SCIENTOMETRICS	Quantitative evaluation of large maps of science
48	Small, H	2006	SCIENTOMETRICS	Tracking and predicting growth areas in science
58	Lucio-Arias, D	2007	SCIENTOMETRICS	Knowledge emergence in scientific communication: from "fullerenes" to "nanotubes"

(continued)

Table 2.9 (continued)

Node	Author	Year	Journal	Title
60	Shibata, N	2007	JASIST	Topological analysis of citation networks to discover the future core articles
67	Bornmann, L	2007	SCIENTOMETRICS	Row-column (RC) association model applied to grant peer review
72	Chavalarias, D	2008	SCIENTOMETRICS	Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study
76	Bettencourt, LMA	2008	SCIENTOMETRICS	Population modeling of the emergence and development of scientific fields
79	Bornmann, L	2008	JOURNAL OF INFORMETRICS	Latent Markov modeling applied to grant peer review
90	Lucio-Arias, D	2008	JASIST	Main-path analysis and path-dependent transitions in HistCite (TM)-based historiograms
102	Borner, K	2009	JOURNAL OF INFORMETRICS	Visual conceptualizations and models of science
103	Chen, CM	2009	JOURNAL OF INFORMETRICS	Towards an explanatory and computational theory of scientific discovery
104	Bettencourt, LMA	2009	JOURNAL OF INFORMETRICS	Scientific discovery and topological transitions in collaboration networks
105	Frenken, K	2009	JOURNAL OF INFORMETRICS	Spatial scientometrics: Towards a cumulative research program
106	Skupin, A	2009	JOURNAL OF INFORMETRICS	Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization
107	Lucio-Arias, D	2009	JOURNAL OF INFORMETRICS	The dynamics of exchanges and references among scientific texts, and the autopoiesis of discursive knowledge

Table 2.10 Documents in Fig. 2.5

Node	Author	Year	Journal	Title
2	Xekalaki, E	2005	SCIENTOMETRICS	Comments on the paper of Shan et al.: The multivariate Waring distribution
7	Egghe, L	2005	JASIST	The power of power laws and an interpretation of Lotkaian informetric systems as self-similar fractals

(continued)

Table 2.10 (continued)

Node	Author	Year	Journal	Title
13	Burrell, QL	2005	SCIENTOMETRICS	The use of the generalized Waring process in modelling informetric data
34	Egghe, L	2006	SCIENTOMETRICS	An informetric model for the Hirsch-index
35	Glanzel, W	2006	SCIENTOMETRICS	On the h-index – A mathematical approach to a new measure of publication activity and citation impact
43	Zitt, M	2006	INFORMATION PROCESSING & MANAGEMENT	Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences
52	Burrell, QL	2007	JOURNAL OF INFORMETRICS	Hirsch's h-index: A stochastic model
56	Egghe, L	2007	JASIST	Dynamic h-index: The Hirsch index in function of time
81	Egghe, L	2008	JASIST	A Model for the Size-Frequency Function of Coauthor Pairs
95	Egghe, L	2008	SCIENTOMETRICS	The mathematical relation between the impact factor and the uncitedness factor

Table 2.11 Documents in Fig. 2.3

Node	Author	Year	Journal	Title
14	Morris, SA	2005	JASIST	Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution
39	Samoylenko, I	2006	JASIST	Visualizing the scientific world and its evolution
50	Kretschmer, H	2007	JOURNAL OF INFORMETRICS	Lotka's distribution and distribution of co-author pairs' frequencies
53	Leydesdorff, L	2007	JASIST	Visualization of the citation impact environments of scientific journals: An online mapping exercise
66	Morris, SA	2007	JASIST	Manifestation of research teams in journal literature: A growth model of papers, coauthorship, weak ties, authors, collaboration, and Lotka's law
68	de Moya-Anegon, F	2007	JASIST	Visualizing the marrow of science

References

Bak P (1996) How nature works: The science of self-organized criticality. Copernicus, New York, NY

Barabási AL (2002) Linked: The new science of networks. Perseus Publishing, Cambridge, MA

Batagelj V (2003) Efficient algorithms for citation network analysis. Arxiv preprint arXiv:cs/0309023

Bettencourt LMA, Kaiser DI, Kaur J (2009) Scientific discovery and topological transitions in collaboration networks. J Informetr 3(3):210–221 (DOI: 10.1016/j.joi.2009.03.001)

Börner K, Sanyal S, Vespignani A (2007) Network science. Ann Rev Inf Sci Technol 41:537–607 (DOI: 10.1002/aris.2007.1440410119)

Boyack KW, Klavans R (2009) The fabric of science. Presentation given at the workshop 'Modelling science dynamics', Amsterdam, 2009, also see the URL: http://scimaps.org/maps/map/weaving_the_fabric_o_119

Bruckner E, Ebeling W, Scharnhorst A (1990) The application of evolution models in scientometrics. Scientometrics 18(1):21–41 (DOI: 10.1007/BF02019160)

Burger M, Bujdoso E (1985) Oscillating chemical reactions as an example of the development of a subfield of science. In: Field RJ, Burger M (eds) Oscillations and traveling waves in chemical systems. Wiley, New York, NY, pp 565–604

Carley KM, Hummon NP, Harty M (1993) Scientific influence: An analysis of the main path structure in the Journal of Conflict Resolution. J Conflict Resolut 14(4):417–447 (DOI: 10.1177/107554709301400406)

Chen C (2006) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci Technol 57(3):359–377 (DOI: 10.1002/asi.20317)

Cronin B (1984) The citation process: The role and significance of citations in scientific communication. Taylor Graham, London

De Bellis N (2009) Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics. Scarecrow, Lanham, MD

Domínguez Reboiras F, Villalba Varneda P, Walter P (eds) (2002) Arbor scientiae: Der Baum des Wissens von Ramon Lull. Akten des Internationalen Kongresses aus Anlaß des 40-jährigen Jubiläums des RAIMUNDUS-LULLUS-INSTITUTS der Universität Freiburg i. Br.. Instrumenta Patristica et Mediaevalia – Research on the Inheritance of Early and Medieval Christianity, vol 42, Subsidia Lulliana, vol 1. Brepols Publishers, Turnhout

Ebeling W, Scharnhorst A (2009) Selbstorganisation und Mobilität von Wissenschaftlern – Modelle für die Dynamik von Problemfeldern und Wissenschaftsgebieten. In: Ebeling W, Parthey H (eds) Selbstorganisation in Wissenschaft und Technik: Wissenschaftsforschung Jahrbuch 2008. Wissenschaftlicher Verlag Berlin, Berlin, pp 9–27, available online at the URL: http://www.wissenschaftsforschung.de/pbuch2008.html

Egghe L (2005) Power laws in the information production process: Lotkaian informetrics. Library and Information Science, vol 5. Elsevier, Amsterdam, available online at the URL: http://www.emeraldinsight.com/books.htm?issn=1876-0562&volume=05

Egghe L, Rousseau R (1990) Introduction to informetrics: Quantitative methods in library, documentation and information science. Elsevier, Amsterdam

Epstein JM, Axtell R (1996) Growing artificial societies: Social science from the bottom up. Complex Adaptive Systems. Brookings Institution Press, Washington, DC; MIT Press, Cambridge, MA

Fronczak P, Fronczak A, Hołyst JA (2007) Publish or perish: Analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. Phys Rev E 75(2):026103, p 9 (DOI: 10.1103/PhysRevE.75.026103), also available as arXiv preprint arXiv:physics/0606190

Garfield E (1979) Citation indexing: Its theory and application in science, technology, and humanities. Information Sciences Series. Wiley, New York, NY

Garfield E, Sher IH, Torpie RJ (1964) The use of citation data in writing the history of science. Technical report, Institute for Scientific Information Inc. Philadelphia, PA. Report of research for the Air Force Office of Scientific Research under contract AF49 (638)-1256, also available online at the URL: http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf

Garfield E, Pudovkin AI, Istomin VS (2003a) Mapping the output of topical searches in the Web of Knowledge and the case of Watson-Crick. Inf Technol Libr 22(4):183–188, available online at the URL: http://www.garfield.library.upenn.edu/papers/casewatsoncrick2003.pdf

Garfield E, Pudovkin AI, Istomin VS (2003b) Why do we need Algorithmic Historiography? J Am Soc Inf Sci Technol 54(5):400–412 (DOI: 10.1002/asi.10226), also available online at the URL: http://www.garfield.library.upenn.edu/papers/jasist54(5)400y2003.pdf

Garfield E, Pudovkin AI, Istomin VS (2002) Algorithmic citation-linked histography – mapping the literature of science. Proc Am Soc Inf Sci Technol 39:14–24 (DOI: 10.1002/meet.1450390102), also available online at the URL: http://www.garfield.library.upenn.edu/papers/asist2002proc.pdf

Gilbert GN, Mulkay MJ (1984) Opening Pandora's box: A sociological analysis of scientists' discourse. Cambridge University Press, Cambridge

Gilbert GN, Troitzsch KG (2005) Simulation for the social scientist, 2nd edn. Open University Press, Maidenhead

Gilbert N (1997) A simulation of the structure of academic science. Sociol Res Online 2(2), available online at the URL: http://www.socresonline.org.uk/2/2/3.html

Glänzel W, Schubert A (1995) Predictive aspects of a stochastic model for citation processes. Inf Process Manag 31(1):69–80 (DOI: 10.1016/0306-4573(95)80007-G)

Goffman W (1966) Mathematical approach to the spread of scientific ideas – the history of mast cell research. Nature 212(5061):449–452 (DOI: 10.1038/212449a0)

Goffman W, Newill VA (1964) Generalization of epidemic theory: An application to the transmission of ideas. Nature 204(4955):225–228 (DOI: 10.1038/204225a0)

Harmon G (2008) Remembering William Goffman: Mathematical information science pioneer. Inf Process Manag 44(4):1634–1647 (DOI: 10.1016/j.ipm.2007.12.004)

Havemann F (2009) Einführung in die Bibliometrie. Gesellschaft für Wissenschaftsforschung, Berlin, available online at the URL: http://edoc.hu-berlin.de/oa/books/reMKADKkid1Wk/PDF/20uf7RZtM6ZJk.pdf

Hummon NP, Doreian P (1989) Connectivity in a citation network: The development of DNA theory. Soc Networks 11(1):39–63 (DOI: 10.1016/0378-8733(89)90017-8)

Krampen G, Becker R, Wahner U, Montada L (2007) On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. Scientometrics 71(2):191–202 (DOI: 10.1007/s11192-007-1659-2)

Kranakis E, Leydesdorff L (1989) Teletraffic conferences: Studying a field of engineering science. Scientometrics 15(5):563–591 (DOI: 10.1007/BF02017071)

Kretschmer H, Kretschmer T (2007) Lotka's distribution and distribution of co-author pairs' frequencies. J Informetr 1(4):308–337 (DOI: 10.1016/j.joi.2007.07.003)

Lambiotte R, Panzarasa P (2009) Communities, knowledge creation, and information diffusion. J Informetr 3(3):180–190 (DOI: 10.1016/j.joi.2009.03.007), also available as arXiv preprint arXiv:0909.2481

Lotka AJ (1911) Die Evolution vom Standpunkte der Physik. Ostwalds Annalen der Naturphilosophie 10:59–74, also available online at the URL: http://digital.slub-dresden.de/id15325443L/65

Lotka AJ (1925) Elements of physical biology. Williams & Wilkins Company, Baltimore, MD, also available online at the URL: http://www.archive.org/details/elementsofphysic017171mbp. Reprinted as: Lotka AJ (1956) Elements of mathematical biology. Dover Books on the Biological Sciences. Dover, New York, NY

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–323

Lucio-Arias D, Leydesdorff L (2008) Main-path analysis and path-dependent transitions in HistCite$^{\text{TM}}$-based historiograms. J Am Soc Inf Sci Technol 59(12):1948–1962 (DOI: 10.1002/asi.v59:12), also available as arXiv preprint arXiv:0911.1454

Lucio-Arias D, Leydesdorff L (2009) The dynamics of exchanges and references among scientific texts, and the autopoiesis of discursive knowledge. J Informetr 3(3):261–271 (DOI: 10.1016/j.joi.2009.03.003), also available as arXiv preprint arXiv:0911.1308

MacRoberts MH, MacRoberts BR (1987) Testing the Ortega hypothesis: Facts and artifacts. Scientometrics 12(5):293–295 (DOI: 10.1007/BF02016665)

MacRoberts MH, MacRoberts BR (1989) Problems of citation analysis: A critical review. J Am Soc Inf Sci 40(5):342–349 (DOI: 10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U)

Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems: From dissipative structures to order through fluctuations. Wiley, New York, NY

Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the Internet: A statistical physics approach. Cambridge University Press, Cambridge (DOI: 10.1017/CBO9780511610905)

Price DJ de Solla (1965) Networks of scientific papers. Science 149(3683):510–515 (DOI: 10.1126/science.149.3683.510). Reprinted in: Price DJ de Solla (1986) Little science, big science . . . and beyond. Columbia University Press, New York, NY, pp 103–118

Price DJ de Solla (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306 (DOI: 10.1002/asi.4630270505), also available online at the URL: http://www.asis.org/Publications/JASIS/Best_Jasist/1976pricejasistarticle.pdf

Pudovkin AI, Garfield E (2002) Algorithmic procedure for finding semantically related journals. J Am Soc Inf Sci Technol 53(13):1113–1119 (DOI: 10.1002/asi.10153), also available online at the URL: http://www.garfield.library.upenn.edu/papers/pudovkinsemanticallyrelatedjournals2002.html

Pyka A, Scharnhorst A (2009) Introduction. Network perspectives on innovations: Innovative networks – network innovation. In: Pyka A, Scharnhorst A (eds) Innovation networks. New approaches in modelling and analyzing. Springer Complexity: Understanding Complex Systems. Springer, Dordrecht, pp 1–16 (DOI: 10.1007/978-3-540-92267-4_1)

Scharnhorst A (1988) Philosophische, methodologische und erkenntnistheoretische Probleme der Anwendung von Instrumentarien der Physik der Selbstorganisation auf quantitative Aspekte der Wissenschaftsentwicklung. Dr. phil. (PhD) Thesis. Humboldt-Universität zu Berlin, Berlin

Scharnhorst A (2003) Complex networks and the web: Insights from nonlinear physics. J Comput Mediat Commun 8(4), available online at the URL: http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html

Scharnhorst A, Garfield E (2010) Tracing scientific influence. Dyn Socio-Econ Syst Int J 2(1):1–33, available online at the URL: http://www.dyses.org.ar/ojs2333/index.php/jdyses/article/view/11, also available as arXiv preprint arXiv:1010.3525

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. Technol Forecast Soc Change 28(2):93–122 (DOI: 10.1016/0040-1625(85)90009-5)

Sterman JD (1992) Response to "On the very idea of a system dynamics model of Kuhnian science". Syst Dyn Rev 8(1):35–42 (DOI: 10.1002/sdr.4260080104)

Tabah AN (1999) Literature dynamics: Studies on growth, diffusion, and epidemics. Ann Rev Inf Sci Technol 34:249–86

van Raan AFJ (2006) Statistical properties of bibliometric indicators: Research group indicator distributions and correlations. J Am Soc Inf Sci Technol 57(3):408–430 (DOI: 10.1002/asi.20284)

Vinkler P (1996) Relationships between the rate of scientific development and citations. The chance for citedness model. Scientometrics 35(3):375–386 (DOI: 10.1007/BF02016908)

von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton, NJ

Wagner-Döbler R (1999) William Goffman's "Mathematical approach to the prediction of scientific discovery" and its application to logic, revisited. Scientometrics 46(3):635–645 (DOI: 10.1007/BF02459617)

Weisberg M, Muldoon R (2009) Epistemic landscapes and the division of cognitive labor. Philos Sci 76(2):225–252 (DOI: 10.1086/644786, stable JSTOR URL: http://www.jstor.org/stable/10.1086/644786)

Wouters P (1999) The citation culture. Ph.D. Thesis. University of Amsterdam, Amsterdam, available online at the URL: http://garfield.library.upenn.edu/wouters/wouters.pdf

Yablonsky AI (1980) On fundamental regularities of the distribution of scientific productivity. Scientometrics 2(1):3–34 (DOI: 10.1007/BF02016597)

Yin L, Kretschmer H, Hanneman RA, Liu Z (2006) The evolution of a citation network topology: The development of the journal Scientometrics. In: Proceedings of the international workshop on webometrics, informetrics and scientometrics & seventh COLLNET meeting on 10-12 May, 2006, Nancy (France). SRDI – INIST – CNRS, Nancy, pp 92–113, available online at the URL: http://hdl.handle.net/10760/7576

AUTHOR QUERIES

AQ1. First author has been considered as corresponding author. Please check.

AQ2. Please provide reference list for Barabási (1999) and Albert (2002), Van Raan (2008a,b)

AQ3. Please check the inserted citation of Table 2.4 is appropriate.

AQ4. Please provide the citation for Tables 2.6 to 2.11

Part II
Exemplary Model Types

Chapter 3

Knowledge Epidemics and Population Dynamics Models for Describing Idea Diffusion

Nikolay K. Vitanov and Marcel R. Ausloos

3.1 Knowledge, Capital, Science Research, and Ideas Diffusion

3.1.1 Knowledge and Capital

Knowledge can be defined as a dynamic framework connected to cognitive structures from which information can be sorted, processed and understood (Howells 2002). Along economics lines of thought (Barro and Sala-I-Martin 2004; Leydesdorff 2006; Dolfsma 2008), knowledge can be treated as one of the "production factors", – i.e., one of the main causes of wealth in modern capitalistic
AQ2 societies (Tables 3.1–3.5).

According to Marshall (Marshall 1920) a "**capital**" is a collection of goods external to the economic agent that can be sold for money and from which an income can be derived. Often, knowledge is parametrized as such a "**human capital**" (Romer 1996, 1994a,b, 2002; Jaffe and Trajtenberg 2002). Walsh (1935) was one pioneer in treating human knowledge as if it was a "capital", in the economic sense; he made an attempt to find measures for this form of "capital". Bourdieu (1986); Coleman (1988), Putnam Putnam (1993), Becker and collaborators have further implanted the concept of such a "human capital" in economic theory (Becker and Murphy 1988; Becker 1996; Stiglitz 1987).

N.K. Vitanov (✉)
AQ1 Institute of Mechanics, Bulgarian Academy of Sciences, Akad. G. Bonchev Str., Bl. 4,
1113 Sofia, Bulgaria
e-mail: vitanov@imbm.bas.bg

M.R. Ausloos
7 rue des Chartreux, B-4122 Plainevaux, Belgium Previously at GRAPES, SUPRATECS,
University of Liege, B-4000 Liege, Belgium
e-mail: marcel.ausloost@ulg.ac.be

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_3,

Table 3.1 Several questions and answers that should guide and supply useful and important information for the reader

10 Important questions raised in this chapter	And their answers in the form of guidance
1. What is the connection between knowledge and capital?	Knowledge is often considered as a form of human capital
2. What happens in the case of knowledge diffusion?	Knowledge is transferred when the subjects interact
3. Should quantitative research be supplemented by qualitative research?	Yes, surely supplemented coordinated joint aims are useful
4. Who are the pioneers of scientometrics?	Alfred Lotka and Derek Price
5. What is the relation between epidemic models and of population dynamics models?	Epidemic models are a particular case of population dynamics models
6. What has to be done if fluctuations strongly influence the system evolution?	Switch from deterministic to stochastic models and think
7. Why are discrete models useful?	Often data is collected for some period of time. Thus, such data is best described by discrete models
8. Around which statistical law are grouped all statistical tools described in the chapter?	Around Lotka law
9. Are all possibly relevant models, presented in this chapter?	NO ! Only an appropriate selection. For more models, consult the literature or ask a specialist
10. What is the strategy followed by the authors of the chapter?	Proceed from simple to more complicated models and from deterministic to stochastic models supplemented by statistical tools

Table 3.2 List of models described in the chapter with comments on their usefulness

Models described in this chapter	Are useful for
Science landscapes	Evaluation of research strategies. Decisions about personal development and promotion
Verhulst Logistic curve	Description of a large class of growth processes
Broadcasting model of technology diffusion	Understanding the influence of mass media on technology diffusion
Word-of-mouth model	Understanding the influence of interpersonal contacts on technology diffusion
Mixed information source model	Understanding the influence of both mass media and interpersonal contacts on technology diffusion
Lotka–Volterra model of innovation diffusion with time lag	Understanding the influence of the time lag between hearing about innovation and its adoption
Price model of knowledge growth with time lag	Modeling the growth of discoveries, inventions, and scientific laws
SIR models of scientific epidemics	Modeling the epidemic stage of scientific idea spreading
SEIR models of scientific epidemics	Extends the SIR model by specifically adding the role of a class of scientists exposed to some scientific idea

Table 3.3 List of models described in the chapter with comments on their usefulness (Continuing Table 3.2)

Models described in this chapter	Are useful for
Discrete model for the change in the number of authors in a scientific field	Modeling and forecasting the evolution in the number of authors and papers in a scientific field
Daley model	Modeling the evolution of a population of papers in a scientific field
Coupled discrete model for populations of scientists and papers	Modeling and forecasting the joint evolution of population of scientists and papers in a research field
Goffman–Newill model for the joint evolution of one scientific field and one of its sub-fields	Epidemic model for the increase of number of scientists from a research field who start work in a sub-field of the scientific field. The model also describes the increase in the number of papers in the research sub-field
Bruckner–Ebeling–Scharnhorst model for the evolution of *n* scientific fields	Understanding the joint evolution of scientific fields in presence of migration of scientists from one field to another field

Table 3.4 List of models described in the chapter with comments on their usefulness (Continuation of Table 3.2)

Models described in this chapter	Are useful for
SI model for the probability of intellectual infection	Modeling the spread of intellectual infection along a scientific network
SEI model for the probability of intellectual infection	Modeling the spread of intellectual infection along a scientific network in the presence of a class of scientists exposed to the intellectual infection
Stochastic evolution model	Modeling the number of scientists in a research subfield as a stochastic variable described by a master equation
Stochastic model of scientific productivity	Modeling the influence of fluctuations in scientific productivity through differential equations for the dynamics of a scientific community
Model of competition between ideologies	Understanding the competition between ideologies with possible migration of believers
Reproduction-transport model	Modeling the change of research field as a migration process

However, the concept of knowledge as a form of capital is an oversimplification. This global-like concept does not account for many properties of knowledge strictly connected to the individual, such as the possibility for different learning paths or different views, multiple levels of interpretation, and different preferences (Davis 2003). In fact, knowledge develops in a quite complex social context, within possibly different frameworks or time scales, and involves "tacit dimensions" (beside the basic space and time dimensions) requiring coding and decoding (Dolfsma 2008).

Table 3.5 List of laws discussed in the chapter with a few words on their usefulness (Continuation of Table 3.2)

Laws described in this chapter	Are useful for
Lotka law	Describing the number distribution of scientists with respect to the number of papers they wrote
Pareto distribution	Writing a continuous version of Lotka law
Zipf law and Zipf–Mandelbrot law	Ranking scientists by the number of papers they wrote
Bradford law	Reflecting the fact that a large number of relevant articles are concentrated in a small number of journals

Key point Nr. 1
Knowledge is much more than a form of capital: it is a dynamic framework connected to cognitive structures from which information can be sorted, processed and understood.

3.1.2 Growth and Exchange of Knowledge

Science policy-makers and scholars have for many decades wished to develop quantitative methods for describing and predicting the initiation and growth of science research (Price 1951, 1971; Foray 2004). Thus, scientometrics has become one of the core research activities in view of constructing science and technology indicators (van Raan 1997).

The accumulation of the knowledge in a country's population arises either from acquiring knowledge from abroad or from internal engines (Nonaka 1994; Nonaka and Konno 1998; Nonaka and Takeuchi 1995; Bernius 2010). The main engines for the production of new knowledge in a country are usually: the public research institutes, the universities and training institutes, the firms, and the individuals (Dahlman 2009). The users of the knowledge are firms, governments, public institutions (such as the national education, health, or security institutions), social organizations, and any concerned individual. The knowledge is transferred from producers to the users by dissemination that is realized by some flow or diffusion of process (Dahlman et al. 2007), sometimes involving physical migration.

Knowledge typically appears at first as purely tacit: *a person "has" an idea* (Saviotti 1999; Cowan and Foray 1997). This tacit knowledge must be codified for further use; after codification, knowledge can be stored in different ways, as in textbooks or digital carriers. It can be transferred from one system to another. In addition to knowledge creation, a system can gain knowledge by knowledge exchange and/or trade.

In knowledge diffusion, the knowledge is transferred while subjects interact (Jaffe 1986; Antonelli 1996; Morone and Taylor 2010). Pioneering studies on knowledge diffusion investigated the patterns through which new technologies are spread in social systems (Rogers 1962; Casetti and Semple 1969). The gain of knowledge due to knowledge diffusion is one of the keys or leads to innovative products and innovations (Kucharavy et al. 2009; Ebeling and Scharnhorst 1985).

Key point Nr. 2
An innovative product or a process is **new** for the group of people who are likely to use it. Innovation is an innovative product or process that has passed the barrier of user adoption. Because of the rejection by the market, many innovative products and processes never become an innovation.

In science, the diffusion of knowledge is mainly connected to the transfer of scientific information by publications. It is accepted that the results of some research become completely scientific when they are published (Ziman 1969). Such a diffusion can also take place at scientific meetings and through oral or other exchanges, sometimes without formal publication of exchanged ideas.[1]

Key point Nr. 3
Scientific communication has specific features. For example, citations are very important in the communication process as they place corresponding research and researchers, mentioned in the scientific literature, in a way similar to the kinship links that tie persons within a tribe. Informal exchanges happening in the process of common work at the time of meetings, workshops, or conferences may accelerate the transfer of scientific information, whence the growth of knowledge.

3.2 Qualitative Research: Historical Remarks

3.2.1 Science Landscapes

Understanding the diffusion of knowledge requires research complementary to mathematical investigations. For example, mathematics cannot indicate why the

---

[1] For example, at Gordon Research Conferences, it is forbidden to take written notes and to quote participant interventions later.

exposure to ideas leads to intellectual epidemics. Yet, mathematics can provide information on the intensity or the duration of some intellectual epidemics.

Qualitative research is all about exploring issues, understanding phenomena, and answering questions (Bryman 1988) without much mathematics. Qualitative research involves empirical research through which the researcher explores relationships using a textual methodology rather than quantitative data. Problems and results in the field of qualitative research on knowledge epidemics will not be discussed in detail here. However, through one example it can be shown how mathematics can create the basis for qualitative research and decision making. This example is connected to the science landscape concepts outlined here below.

The idea of science landscapes has some similarity with the work of Wright (1932) in biology who proposed that the fitness landscape evolution can be treated as optimization process based on the roles of mutation, inbreeding, crossbreeding, and selection. The science landscape idea was developed by Small (1997, 1998), as well as by Noyons and Van Raan (1998). In this framework, Scharnhorst (1998, 2001) proposed an approach for the analysis of scientific landscapes, named "geometrically oriented evolution theory".

Key point Nr. 4

The concept of science landscape is rather simple: Describe the corresponding field of science or technology through a function of parameters such as height, weight, size, technical data, etc. Then a virtual knowledge landscape can be constructed from empirical data in order to visualize and understand innovation and to optimize various processes in science and technology.

As an illustration at this level, consider that a mathematical example of a technological landscape can be given by a function $C = C(S, v)$, where C is the cost for developing a new airplane, and where S and v represent the size and velocity of the airplane.

Consider two examples concerning the use of science landscapes for evaluation purposes:

(1) Science landscape approach as a method for evaluating national research strategies

For example, national science systems can be considered as made of researchers who compete for scientific results, and subsidies, following optimal research strategies. The efforts of every country become visible, comparable and measurable by means of appropriate functions or landscapes: e.g., the number of publications. The aggregate research strategies of a country can thereby be represented by the distribution of publications in the various scientific disciplines. In so doing, within a two-dimensional space,[2] different countries correspond to different landscapes. Various political discussions can follow and evolution strategies can be invented thereafter.

Notice that the dynamics of self-organized structures in complex systems can be understood as the result of a search for optimal solutions to a certain problem. Therefore, such a comment shows how rather strict mathematical approaches, not disregarding simulation methods, can be congruent to qualitative questions.

(2) Scientific citations as landscapes for individual evaluation
Scientific citations can serve for constructing landscapes. Indeed, citations have a key position in the retrieval and valuation of information in scientific communication systems (Scharnhorst 1998; Egghe 1998; Egghe and Rousseau 1990). This position is based on the objective nature of the citations as components of a global information system, as represented by the Science Citation Index. A landscape function based on citations can be defined in various ways. It can take into account self-citations (Hellsten et al. 2006, 2007a,b; Ausloos et al. 2008), or time-dependent quantitative measures (Hirsch 2005; Soler 2007; Burrell 2007).

Key point Nr. 5
Citation landscapes become important elements of a science policy (e.g., in personnel management decisions), thereby influencing individual scientific careers, evaluation of research institutes, and investment strategies.

3.2.2 Lotka and Price: Pioneers of Scientometrics

Alfred Lotka, one of the modern founders of population dynamics studies, was also an excellent statistician. He discovered (Lotka 1926) a distribution for the number of authors n_r as a function of the number of published papers r, – i.e., $n_r = n_1/r^2$.

However, Derek Price, a physicist, set the mathematical basis in the field of measuring scientific research in recent times (Price 1963; Price and Gürsey 1975; Price 1961). He proposed a model of scientific growth connecting science and time. In the first version of the model, the size of science was measured by the number of journals founded in the course of a number of years. Later, instead of the number of journals, the number of published papers was used as the measure of scientific growth. Price and other authors (Price and Gürsey 1975; Price 1961; Gilbert 1978) considered also different indicators of scientific growth, such as the number of authors, funds, dissertation production, citations, or the number of scientific books.

---

[2] For example, take the scientific disciplines and the number of publications as axes.

In addition to the deterministic approach initiated by Price, the statistical approach to the study of scientific information developed rapidly and nowadays is still an important tool in scientometrics (Chung and Cox 1990; Kealey 2000). More discussion on the statistical approach will be given in sect. 3.6 of this chapter.

Key point Nr. 6
Price distinguished three stages in the growth of knowledge: **(a)** a preliminary phase with small increments; **(b)** a phase of exponential growth; **(c)** a saturation stage. The stage (c) must be reached sooner or later after the new ideas and opportunities are exhausted; the growth slows down until a new trend emerges and gives rise to a new growth stage. According to Price, the curve of this growth is a S-shaped logistic curve.

3.2.3 Population Dynamics and Epidemic Models of the Diffusion of Knowledge

Population dynamics is the branch of life sciences that studies short- and long-term changes in the size and age composition of populations, and how the biological and environmental processes influence those changes. In the past, most models for biological population dynamics have been of interest only in mathematical biology (Murray 1989; Edelstein-Keshet 1988). Today, these models are adapted and applied in many more areas of science (Dietz 1967; Dodd 1958). Here below, models of knowledge dynamics will be of interest as bases of epidemic models. Such models are nowadays used because some stages of idea spreading processes within a population (e.g, of scientists), possess properties like those of epidemics.

The mathematical modeling of epidemic processes has attracted much attention since the spread of infectious diseases has always been of great concern and considered to be a threat to public health (Anderson and May 1982; Brauer and Castillo-Chavez 2001; Ma and Li 2009). In the history of science and society, many examples of ideas spreading seem to occur in a way similar to the spread of epidemics. Examples of the former field pertain to the ideas of Newton on mechanics and the passion for “High Critical Temperature Superconductivity” at the end of the twentieth century. Examples of the latter field are the spreading of ideas from Moses or Buddha (Goffman 1966), or discussions based on the Kermack–McKendrick model (Kermack and McKendrick 1927) for the epidemic stages of revolutions or drug spreading (Epstein 1997).

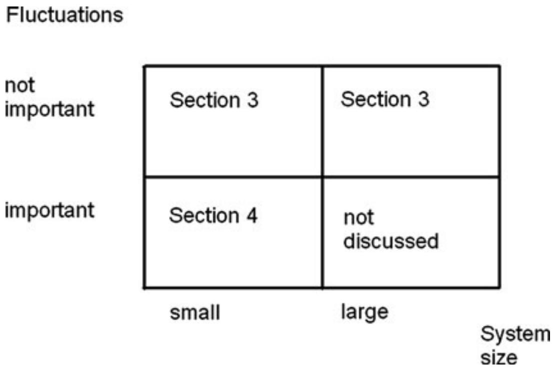
Epidemic models belong to a more general class of Lotka–Volterra models used in research on systems in the fields of biological population dynamics, social dynamics, and economics. The models can also be used for describing processes connected to the spread of knowledge, ideas and innovations (see Fig. 3.1). Two examples are the model of innovation in established organizations (Castiaux 2007) and the Lotka–Volterra model for forecasting emerging technologies and the growth of knowledge (Kucharavy et al. 2009). In social dynamics, the Lanchester model of war between two armies can be mentioned, a model which in the case of reinforcements coincides with the Lotka–Volterra–Gause model for competition between two species (Gause 1935). Solomon and Richmond (2001, 2002) applied a Lotka–Volterra model to financial markets, while the model for the trap of extinction can be applied to economic subjects (Vitanov et al. 2006). Applications to chaotic pairwise competition among political parties (Dimitrova and Vitanov 2004) could also be mentioned.

To start the discussion of population dynamics models as applied to the growth of scientific knowledge with special emphasis on epidemic models, two kinds of models can be discussed (Fig. 3.2): **(1) deterministic models**, see Sect. 3.3, appropriate for large and small populations where the fluctuations are not drastically important, **(2) stochastic models**, see Sect. 3.4, appropriate for small populations. In the latter case the intrinsic randomness appears much more relevant than in the former case. Stochastic models for large populations will not be discussed. The reason for this is that such models usually consist of many stochastic differential equations, whence their evolution can be investigated only numerically.

Finally, let us mention that the knowledge diffusion is closely connected to the structure and properties of the social network where the diffusion happens. This is a new and very promising research area. For example, a combination can be made

Fig. 3.1 Relation among epidemic models, Lotka–Volterra models, and population dynamics models

Fig. 3.2 Relationships between system size, influence of fluctuations, and discussed classes of models



between the theory of information diffusion and the theory of complex networks (Boccaletti et al. 2006). For more information about the relation between networks and knowledge, see the following chapters of the book.

3.3 Deterministic Models

Below, 13 selected deterministic models (see Fig. 3.3) are discussed. The emphasis is on models that can be used for describing the epidemic stage of the diffusion of ideas, knowledge, and technologies.

3.3.1 Logistic Curve and Its Generalizations

In a number of cases, the natural growth of autonomous systems in competition can be described by the logistic equation and the logistic curve (S-curve) (Meyer 1994).

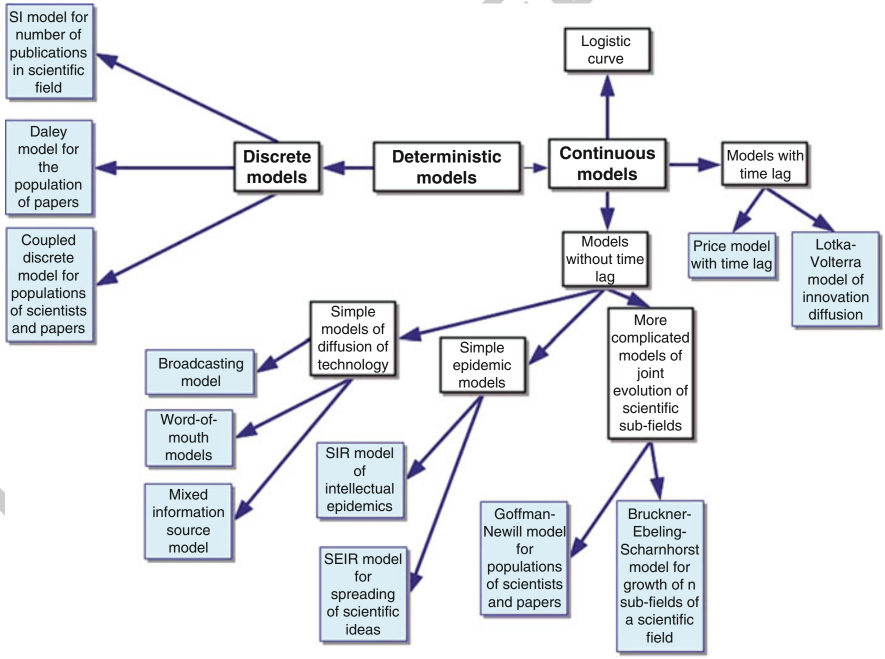


Fig. 3.3 Discrete (3) and continuous (10) models discussed in the chapter. Two continuous models account for the influence of time lag, three models are simple models of technological diffusion. Two models are simple epidemic models and two models are more complicated models. In addition, the basic logistic curve is discussed

In order to describe trajectories of growth or decline in socio-technical systems, one generally applies a three-parameter logistic curve:

$$N(t) = \frac{K}{1 + \exp[-\alpha t - \beta]} \tag{3.1}$$

where $N(t)$ is the number of units in the species or growing variable to study; K is the asymptotic limit of growth; α is the growth rate which specifies the "width" of the S-curve for $N(t)$; and β specifies the time t_m when the curve reaches the midpoint of the growth trajectory, such that $N(t_m) = 0.5\ K$. The three parameters, K, α, and β, are usually obtained after fitting some data (Meade and Islam 1995). It is well known that many cases of epidemic growth can be described by parts of an appropriate S-curve. As an example, recall that the S-curve was also used for describing technological substitution (Rogers 1962; Mansfield 1961; Modis 2007), ca. 60 years ago.

However, different interaction schemes can generate different growth patterns for whatever system species are under consideration (Modis 2003). Not every interaction scheme leads to a logistic growth (Ausloos 2010). The evolution of systems in such regimes may be described by more complex curves, such as a combination of two or more simple three-parameter functions (Meyer 1994; Meyer et al. 1999).

3.3.2 Simple Epidemic and Lotka–Volterra Models of Technology Diffusion

As recalled here above, the simplest epidemic models could be used for describing technology diffusion, like considering two populations/species: adopters and non-adopters of some technology. Such models can be put into two basic classes: either broadcasting (Fig. 3.4) or word-of-mouth models (Fig. 3.5). In the broadcasting models, the source of knowledge about the existence and/or characteristics of the new technology is external and reaches all possible adopters in the same way. In the word-of-mouth models, the knowledge is diffused by means of personal interactions.

(1) The broadcasting model (Fig. 3.4)

Let us consider a population of K potential adopters of the new technology and let each adopter switch to the new technology as soon as he/she hears about its existence (immediate infection through broadcasting). The probability that at time t a new subject will adopt the new technology is characterized by a coefficient of diffusion $\kappa(t)$ which might or might not be a function of the number of previous adopters. In the broadcasting model $\kappa(t) = a$ with $(0 < a < 1)$; this is considered to be a measure of the infection probability.

Let $N(t)$ be the number of adopters at time t. The increase in adopters for each period is equal to the probability of being infected, multiplied by the current

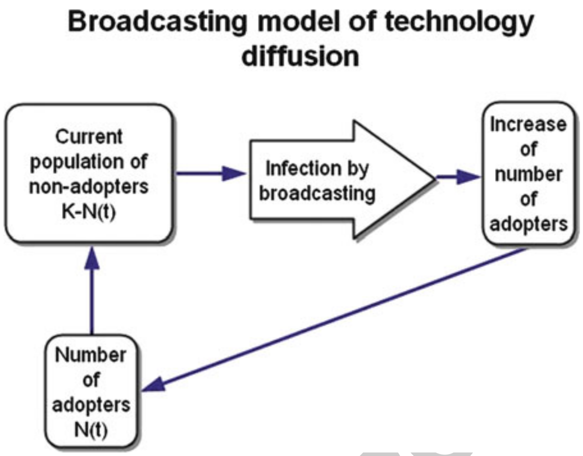


Fig. 3.4 Schematic representation of a broadcasting model of technology diffusion. The number of adopters of technology increases by mass media influence

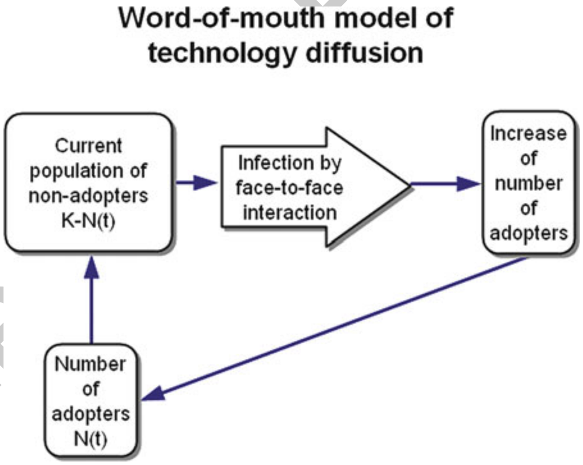


Fig. 3.5 Schematic representation of a word-of-mouth model of technology diffusion. The number of adopters of technology increases by interpersonal interactions

population of non-adopters (Mahajan and Peterson 1985). The rate of diffusion at time t is

$$\frac{dN}{dt} = a[K - N(t)]. \tag{3.2}$$

The integration of (3.2) leads to the number of adopters: i.e.,

$$N(t) = K[1 - \exp(-at)]. \tag{3.3}$$

$N(t)$ is described by a decaying exponential curve.

(2) Word-of-mouth model (Fig. 3.5)

In many cases, however, the technology adoption timing is at least an order of magnitude slower than the time it takes for information spreading (Geroski 2000). This requires another modelization than in **(1)**: the word-of-mouth diffusion model. Its basic assumption is that knowledge diffuses by means of face-to-face interactions. Then the probability of receiving the relevant knowledge needed to adopt the new technology is a positive function of current users $N(t)$. Let the coefficient of diffusion $\kappa(t)$ be $bN(t)$ with $b > 0$. The rate of diffusion at time t is

$$\frac{dN}{dt} = b\ N(t)\ [K - N(t)]\ . \tag{3.4}$$

Then

$$N(t) = \frac{K}{1 + \left(\dfrac{K - N_0}{N_0}\right) e^{-bK(t-t_0)}} \tag{3.5}$$

where $N_0 = N(t = t_0)$. $N(t)$ is described by an S-shaped curve.

A constraint exists in the word-of-mouth model: it explains the diffusion of an innovation not from the date of its invention but from the date when some number, $N(t) > 0$, of early users have begun using it.

(3) Mixed information source model (Fig. 3.6)

In the mixed information source model, existing non-adopters are subject to two sources of information (Fig. 3.6). The coefficient of diffusion is supposed to look like $a + bN(t)$. The model evolution equation becomes

$$\frac{dN}{dt} = (a + bN(t))\ [K - N(t)]. \tag{3.6}$$

The result of (3.6) is a (generalized) logistic curve whose shape is determined by a and b (Mahajan and Peterson 1985).

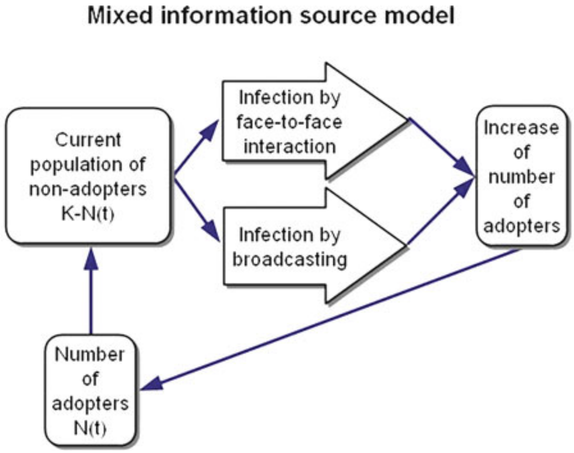


Fig. 3.6 Schematic representation of mixed information source model. The number of adopters increases by mass media influence and interpersonal contacts

(4) Time lag Lotka–Volterra model of innovation diffusion (Fig. 3.7)

Let it be again assumed that the diffusion of innovation in a society is accounted for by a combination of two processes: a mass-mediated process and a process connected to interpersonal (word-of-mouth) contacts. Let $N(t)$ be the number of potential adopters. Some of the potential adopters adopt the innovation and become real adopters. The equation for the he rate of growth of the real adopters $n(t)$, in absence of time lag, is

$$\frac{dn(t)}{dt} = \alpha[N(t) - n(t)] + \beta n(t)[N(t) - n(t)] - \mu n(t), \tag{3.7}$$

where α denotes the degree of external influence such as mass media, β accounts for the degree of internal influence by interpersonal contact between adopters and the remaining population; μ is a parameter characterizing the decline in the number of adopters because of technology rejection for whatever reason.

A basic limitation in most models of innovation diffusion has been the assumption of instantaneous acceptance of the new innovation by a potential adopter (Mahajan and Peterson 1985; Bartholomew 1982). Often, in reality, there is a finite time lag between the moment when a potential adopter hears about a new innovation and the time of adoption. Such time lags usually are continuously distributed (May 1974; Lal et al. 1988).

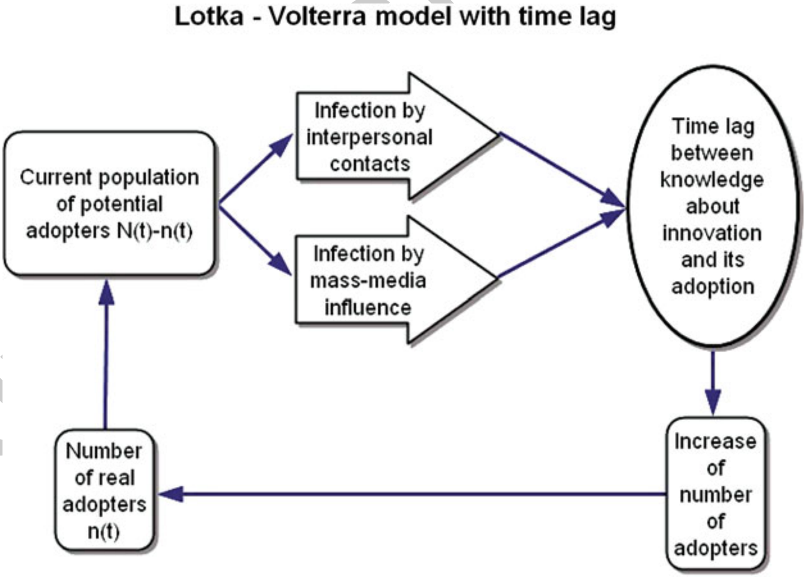


Fig. 3.7 Schematic representation of a Lotka–Volterra model with time lag. The model accounts for the time lag between hearing about innovation and its adoption

this figure will be printed in b/w

The time lag between the knowledge about the innovation and its adoption can be captured by a distributed time lag approach in which the effects of time delays are expressed as a weighted response over a finite time interval through appropriately chosen memory kernels (Karmeshu 1982) (see Fig. 3.7). Whence (3.7) becomes

$$\frac{dn(t)}{dt} = \alpha \int_0^t d\tau \; K_1^*(t-\tau)\,[N(\tau)-n(\tau)] + \beta \int_0^t d\tau \; K_2^*(t-\tau) n(\tau)[N(\tau)-n(\tau)] - \mu \int_0^t d\tau \; K_3^*(t-\tau) n(\tau). \tag{3.8}$$

Equation (3.8) reduces to (3.7) when the memory kernels $K_i^*(t)$ $(i = 1, 2, 3)$ are replaced by delta functions.

Two generic types of kernels are usually considered (Lal et al. 1988):

$$K^*(t) = \nu \; e^{-\nu t} \tag{3.9}$$

$$K^*(t) = \nu^2 t \; e^{-\nu t} \;, \tag{3.10}$$

in which ν^{-1} is some characteristic time scale of the system.

The number of potential adopters $N(t)$ changes over time. Several possible functional forms of $N(t)$ are used (Sharif and Ramanathan 1981):

$$N(t) = N_0(1+at); \qquad N_0 > 0, a > 0 \tag{3.11}$$

$$N(t) = N_0 \exp[gt]; \quad N_0 > 0, g > 0 \tag{3.12}$$

$$N(t) = \frac{b}{1+d\exp(-ct)}; \qquad b > 0, d > 0, c > 0 \tag{3.13}$$

$$N(t) = b - q\exp(-rt); \qquad b > 0, q > 0, r > 0. \tag{3.14}$$

Equation (3.12) represents an approximation for short- and medium-term forecasting since for t large, $N(t)$ grows without bound, as in Keynes (1930). Equations (3.13) and (3.14) are useful in long-term forecasting as $N(t)$ has an upper limit. Such forms for $N(t)$ are valid within a deterministic framework.

However, a stochastic framework (see below) is more appropriate when the carrying capacity $N(t)$ is governed by some stochastic process, as when the influence of socioeconomic and natural factors are subject to "random" or hardly explainable fluctuations. In such systems, $N(t)$ can be time-dependent: for example, $N(t) \sim N_0(1 + \epsilon\cos(\omega t))$ where $\epsilon << 1$ and the periodicity takes into account the influence of some (strong) cyclic economic factors. In presence of a strong stochastic component, $N(t)$ can be stochastic: $N(t) = N_0 + \xi(t)$, where the noisy component is $\xi(t)$ and N_0 is the average value of the so-called carrying capacity (Odum 1959).

Key point Nr. 7
Time lags between observations and decisions lead to complicated dynamics. Perform some preliminary careful analysis of system behavior based on time lags before making a decision.

3.3.3 Price Model of Knowledge Growth: Cycles of Growth of Knowledge

The Price evolution model of scientific growth ignited intensive research (Fernandez-Camo et al. 2004; Szydlowski and Krawiez 2001) (see Fig. 3.8). This model is in fact a dialectical addition to Kuhn's idea (Kuhn 1962) about the revolutionary nature of science processes: after some period of evolutionary

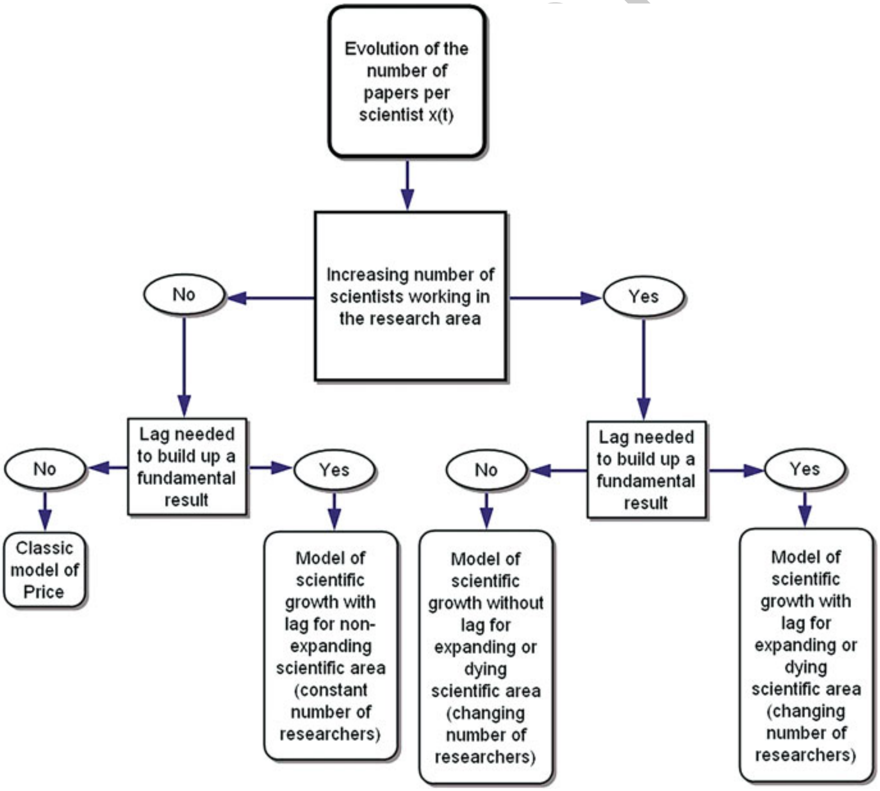


Fig. 3.8 Diagram of relationships between Price model and its modifications. The presence of time lags can lead to much complication in the evolution dynamics of a scientific field

growth, a scientific revolution occurs. Price considered the exponential growth as a disease that retards the growth of stable science, producing narrower and less flexible specialists.

Key point Nr. 8
An interesting result of the research of Price can be read as follows:
if a government wants to double the usefulness of science, it has to multiply by about eight the gross number of workers and the total expenditure of manpower and national income.

The unreserved application of the Price model faces several difficulties:

- Many scientific products which seem to be new are not really new
- Creativity and innovation can be confused (Plesk 1997; Amabile et al. 1996)
- Creative papers with new ideas and results have the same importance as trivial duplications (Magyari-Beck 1984)
- Two things are omitted:
  - Quality (whatever that means, but it is an economic notion) of research
  - The cost or measure of complexity.

In answer to this, Price formulated the hypothesis that one should be studying only the growth of **important** discoveries, inventions, and scientific laws, rather than both important and trivial things. In so doing, one might expect that any of such studied growth will follow the same pattern.

A generalized version of the Price model for the growth of a scientific field (Szydlowski and Krawiez 2009; Price 1956) is based on the following assumptions: (a) the growth is measured by the number of important publications appearing at a given time; (b) the growth has a continuous character, though a finite time period $T =$ const is needed to build up a result of the fundamental character; (c) the interactions between various scientific fields are neglected. If, in addition, the number of scientists publishing results in this field is constant, then the rate of scientific growth is proportional to the number of important publications at time t minus the time period T required to build up a fundamental result. The model equation is

$$\frac{dx}{dt} = \alpha x(t - T), \tag{3.15}$$

where α is a constant. The initial condition $x(t) = \phi(t)$ is defined on the interval $[-T, 0]$.

Let the population of scientists be varying and consider the evolution of the average number of papers per scientist. In general, instead of the linear right-hand side (3.15), a non-linear model can be used:

$$\frac{dx}{dt} = f(x(t - T), x(t)), \tag{3.16}$$

where $f(t-T)$ is a homogeneous function of degree one. The simplest form of such a function is a linear function. Let $n(t)$ represent the rate of growth of the population of scientists and write $L(t) = \exp[n(t) \; t]$. For simplicity, let the population of scientists grow at the constant rate $n = \frac{1}{L}\frac{dL}{dt}$ and let $z = x/L$. Then the evolution of the number of papers written by a scientist has the form

$$\frac{dz}{dt} = \alpha z(t-T) - nz(t). \tag{3.17}$$

If $n = 0$ and $T = 0$, the Price model of exponential growth is recovered. Equation (3.17) is linear, but a cyclic behavior may appear because of the feedback between the delayed and non-delayed terms.

3.3.4 *Models Based on Three or Four Populations: Discrete Models*

(1) SIR (Susceptible-Infected-Removed) model (Fig. 3.9)
In 1927, Kermack and McKendrick (1927) created a model in which they considered a fixed population with only three compartments: $S(t)$, the susceptibles; $I(t)$, the infected; $R(t)$, the recovered, or removed.

Following this idea, Goffman (1966); Goffman and Newill (1964) considered the stages of fast growth of scientific research in a scientific field as "intellectual epidemics" and developed the corresponding scientific research epidemic stage based on three classes of population: (i) the susceptibles S who can become infectives when in contact with infectious material (the ideas); (ii) the infectives I who host the infectious material; and (iii) the recovered R who are removed from the epidemics for different reasons (Fig. 3.9).

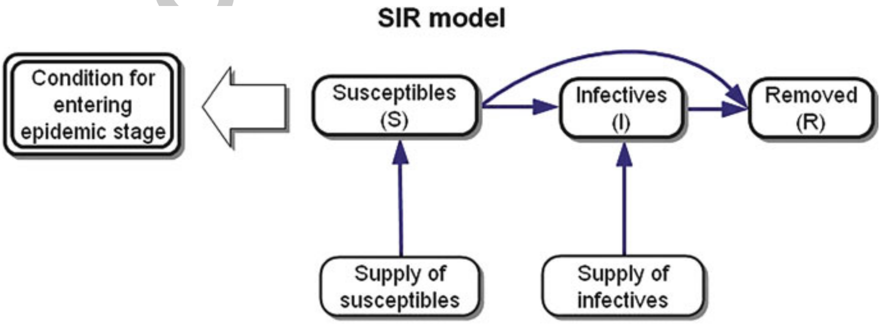


Fig. 3.9 SIR (susceptibles S, infectives I, recovered R) model of intellectual infection with influxes of susceptibles and infectives to the corresponding scientific ideas

The epidemic stage is controlled by the system of differential equations

$$\frac{dS}{dt} = -\beta S I - \delta S + \mu, \tag{3.18}$$

$$\frac{dI}{dt} = \beta S I - \gamma I + \nu, \tag{3.19}$$

$$\frac{dR}{dt} = \delta S + \gamma I \tag{3.20}$$

where μ and ν are the rates at which the new supply of susceptibles and infectives enter the population. A necessary condition for the process to enter the epidemic state is $\frac{dI}{dt} > 0$. Then

$$S > \frac{\gamma - \nu/I}{\beta} = \rho \tag{3.21}$$

is the threshold density of susceptibles, i.e., no epidemics can develop from time t_0 unless S_0, the number of susceptibles at that time, exceeds the threshold ρ: the epidemic state cannot be maintained over some time interval unless the number of susceptibles is larger than ρ through that interval of time. As I increases, ν/I converges to 0 and ρ converges rapidly to γ/β.

In Goffman (1966), Goffman evaluated the rate of change of infectives $\Delta I/\Delta t$. From the system equations, it is difficult to determine $I(t)$. Yet in the epidemic stage, the behaviour of $I(t)$ is exponential. For small t close to t_0, $I(t)$ can be expanded into a power series: $I(t) = C_0 + C_1 t + C_2 t^2 + \dots C_n t^n + \dots$ such that the approximate rate of $\Delta I/\Delta t$ can be obtained. On the basis of this rate and the raw data, the development and peak of some research activity can be predicted, – under the assumption that the research is in an epidemic stage.

(2) SEIR model for the spreading of scientific ideas (Fig. 3.10)
The SIR epidemic models can be further refined by introducing a fourth class, E, i.e., persons exposed to the corresponding scientific ideas (Fig. 3.10). Such models are discussed in Bettencourt et al. (2008, 2006); they belong to the class of so-called SEIR epidemic models. One typical model goes as follows

$$\frac{dS}{dt} = \lambda N - \frac{\beta S I}{N}; \qquad \frac{dE}{dt} = \frac{\beta S I}{N} - \kappa E - \frac{\rho E I}{N}; \tag{3.22}$$

$$\frac{dI}{dt} = \kappa E + \frac{\rho E I}{N} - \gamma I; \qquad \frac{dR}{dt} = \gamma I \tag{3.23}$$

where $S(t)$ is the size of the susceptible population at time t, $E(t)$ is the size of the exposed class, $I(t)$ is the size of the infected class. These individuals have adopted the new scientific idea in their publications. Finally, $R(t)$ is the size of the population of recovered scientists, i.e., those who no longer publish on the topic. The size of the entire population is: $N = S + E + I + R$. An exit term is assumed to be very small, and because of this, t is included in the recovered class. N grows exponentially with

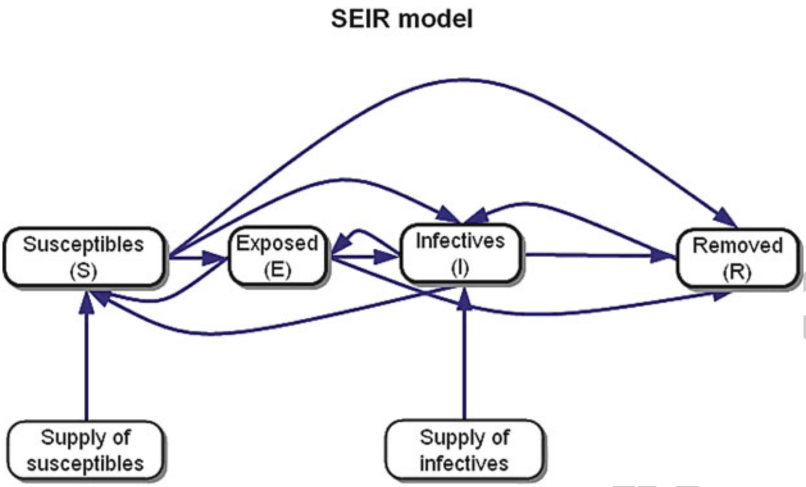


Fig. 3.10 SEIR model of intellectual infection with influxes of susceptibles and infectives to the corresponding scientific ideas, thus extending the SIR model by including a class of scientists exposed (E) to the specific scientific ideas

rate λ. The parameters of the model are: β, the probability and effectiveness of a contact with an adopter; $1/\kappa$, the standard latency time, (in other words, the average duration of time after one has been exposed but before one includes the new idea in one's own publication); $1/\gamma$, the duration of the infectious period, thus how long one publishes on the topic and teaches others; ρ, the probability that an exposed person has multiple effective contacts with other adopters.

This simple model can incorporate a wide range of behaviors. For many values of the parameters λ,β, κ, γ and ρ, the infected class grows as a logistic curve. For large values of the contact rate β or recruitment λ, $I(t)$ grows nearly linearly, as indeed has been found empirically for some research fields (Bettencourt et al. 2008).

Key point Nr. 9
Epidemic models are the best suited for describing the **expansion stage** of a process growth.

(3) SI discrete model for the change in the number of authors in a scientific field (Fig. 3.11)

With the goal of predicting the spreading out of scientific objects (such as theories or methods), Nowakowska (1973) discussed several epidemic discrete models for predicting changes in the number of publications and authors in a given scientific field. With respect to the publications, the main assumption of the models is that the number of publications in the next period of time (say, 1 year) will depend: (i) on the

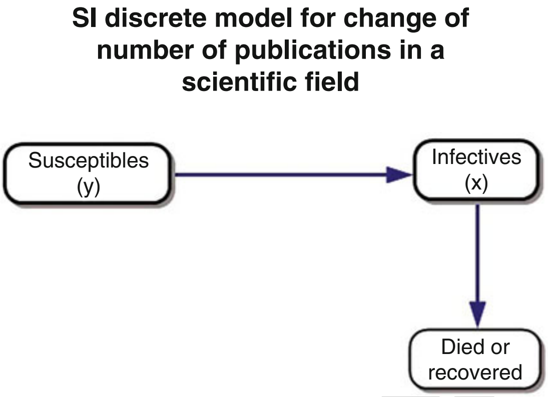


Fig. 3.11 Schema of a discrete SI evolution model of the number of authors of scientific papers. The model takes into account that several scientists stop their work in a scientific field; it can be due to different reasons as for example death or losing interest in particular questions

number of papers which recently appeared, and (ii) on the degree at which the subject has been exhausted. The numbers of publications appearing in successive periods of time should first increase, then would reach a maximum, and as the problem becomes more and more exhausted, the number of publications would decrease.

Let it be assumed (Fig. 3.11) that if at a certain moment t the epidemics state is (x_t, y_t) (x_t is the number of infectives (authors who write papers on the corresponding research problems), y_t is the number of susceptibles), then for a sufficiently short time interval Δt, one may expect that the number of infectives $x_{t+\Delta t}$ will be equal to $x_t - ax_t \Delta t + bx_t y_t \Delta t$, while the number of susceptibles $y_{t+\Delta t}$ will be equal to $y_t - bx_t y_t \Delta t$; a and b being appropriate constants. Let the expected number of individuals who either die or recover, during the interval $(t, t + \Delta t)$, be $ax_t \Delta t$, and let $bx_t y_t \Delta t$ be the expected number of new infections. The equations of this model are:

$$x_{t+\Delta t} = ax_t - ax_t \Delta t + bx_t y_t \Delta t \tag{3.24}$$

$$y_{t+\Delta t} = y_t - bx_t y_t \Delta t. \tag{3.25}$$

Note here that such discrete models are useful for the analysis of realistic situations where the values of the quantities are available at selected moments (every month, every year, etc.).

(4) Daley discrete model for the population of papers (Fig. 3.12)

Daley (1967) investigated the spread of news as follows: individuals who have not heard the news are susceptible and those who heard the news are infective. Recovery is not possible, as it is assumed that the individuals have perfect memory and never forget. The Daley model can be applied also to the population of papers (Nowakowska 1973) (see Fig. 3.12). For $\Delta t = 1$ (year), the Daley model equation reads

$$x_{t+1} = bx_t \left(N - \sum_{i=1}^{t} x_i \right) \tag{3.26}$$

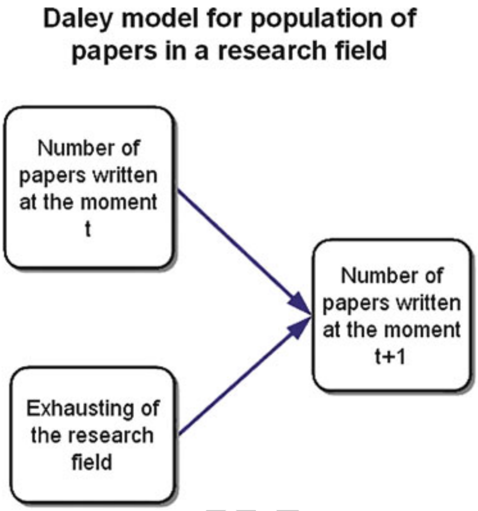


Fig. 3.12 Daley model for evolution of population of papers on problems in a scientific field. The exhausting of the scientific field is taken into account

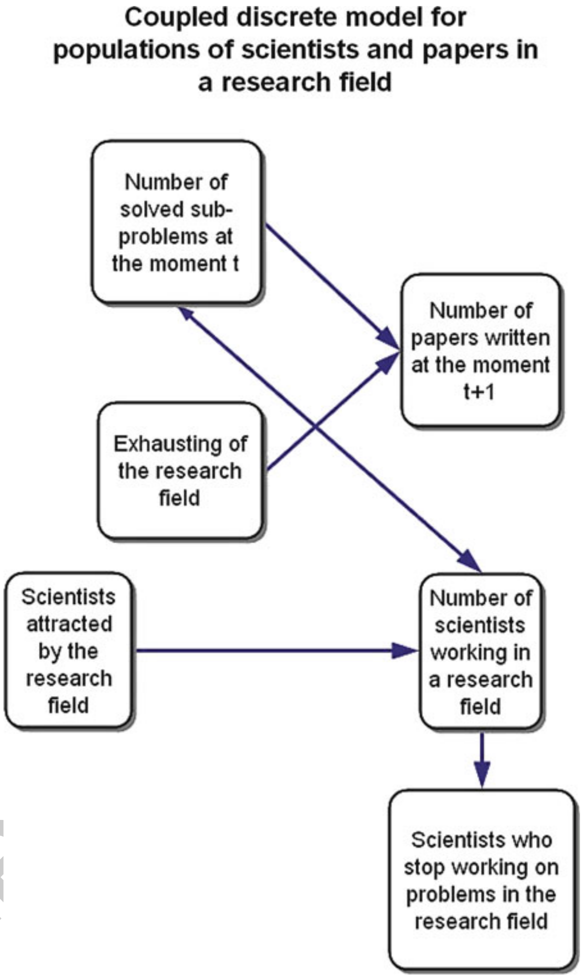
where x_1, x_2 .... are the numbers of papers on the subject which appear in successive periods of time, b and N being parameters. The expected number x_{t+1} of papers in year $t+1$ is proportional to the number x_t of papers which appeared in year t, and to the number $N - x_1 - x_2 \cdots - x_t = N - \sum_{i=1}^{t} x_i$. N is the number of papers which have to appear in order to exhaust the problem: the problem under consideration may be partitioned into N sub-problems, such that solving any of them is worth a separate publication; these subproblems are solved successively by the scientists. The b and N parameters may be estimated by the method of least squares, e.g. from a given empirical histogram. A parameter characterizing the initial growth dynamics in the number of publications can also be introduced: $\tau = bN$. Therefore, (3.26) can be used for short-time prediction, even when the corresponding research field is in the epidemic stage of its evolution.

(5) Discrete model coupling the populations of scientists and papers (Fig. 3.13) A discrete model coupling the populations of scientists and papers can be considered (Fig. 3.13); it depends on four parameters: N, a, b and c. N as above denotes the number of sub-problems of the given problem; a is the probability that a scientist working on the subject in a given year abandons research on the subject for whatever reasons; b is the probability of obtaining a solution to a given subproblem by one scientist during one year of research; c denotes the coefficient of attractiveness of the subject. The basic variables of the model are: u_t, the number of scientists working on the subject in year t, and x_t, the number of publications on the subject which appear in year t.

The model equations are

$$u_{t+1} = (1-a)u_t + cx_t \tag{3.27}$$

Fig. 3.13 Discrete model for the joint evolution of populations of scientists and papers. The attractiveness of the field, the exhaustion of the field, and the possibility for declining interest for working in the scientific field are taken into account through adequate rate parameters



$$x_{t+1} = [1 - (1-b)^{u_t}] \left(N - \sum_{i=1}^{t} x_i \right). \tag{3.28}$$

The equation for the number u_{t+1} of scientists working on the subject in year $t+1$ tells that in year $t+1$, the expected number of scientists working on the subject will be the number of scientists working on the subject in year t, u_t, minus the expected number of scientists who stopped working on the subject, au_t, plus the expected number of scientists, cx_t, who became attracted to the problem by reading papers which appeared in year t. The equation expressing the number of publications in year $t+1$ tells us that x_{t+1} equals the number of subproblems that were solved in the year t. The probability that a given subproblem will be solved in year t by a given scientist equals b. Then the probability of the opposite event, i.e. a given scientist will not solve a particular problem, equals $1 - b$. As there are u_t scientists working on the subject in year t, the probability that a given subproblem will not be solved by any of them is $(1 - b)^{u_t}$. Consequently, the probability that a given subproblem will be solved in year t (by any of the u_t scientists working on the subject) is equal to $1 - (1 - b)^{u_t}$. Next, in year t there remained $N - \sum_{i=1}^{t} x_i$ subproblems to be solved. The expected number of subproblems solved in year t is equal to the product which gives the right-hand side of (3.28).

N.B. It is assumed, that the waiting time for publishing of the paper is one year. A more realistic picture would be to assume that the unit of time is not 1 year, but 2 years, or that the publication has some other time delay.

Key point Nr. 10
In many cases, the data is available as one value per week, or one value per month, or one value per 3 months, etc. For modeling and subsequent short-range forecasting, so-called discrete (time) models are thus very appropriate.

3.3.5 Continuous Models of the Joint Evolution of Scientific Sub-Systems

(1) Coupled continuous model for the populations of scientists and papers: Goffman–Newill model
The Goffman–Newill model (Goffman and Newill 1964) (Fig. 3.14) is based on the idea that the spreading process within a population can be studied on the basis of the literature produced by the members of that population. There is a transfer of infectious materials (ideas) between humans by means of an intermediate host (a written article). Let a scientific field be F and SF a sub-field of F. Let the number of scientists writing papers in the field F at t_0 be N_0 and the number of scientists writing papers in SF at t_0 (the number of infectives) be I_0. Thus, $S_0 = N_0 - I_0$ is the number of susceptibles; there is no removal at t_0, but there is removal $R(t)$ at later times t. The number of papers produced on F at t_0 is N'_0 and the number of papers produced in SF at this time is I'_0. The process of intellectual infection is as follows: (a) a member of F is infected by a paper from I'; (b) after some latency period, this infected member produces 'infected' papers in N', i.e. the infected member produces a paper in the subfield SF citing a paper from I'; (c) this 'infected' paper may infect other scientists from F and its sub-fields, such that the intellectual infection spreads from SF to the other sub-fields of F.

Let β be the rate at which the susceptibles from class S become 'intellectually infected' from class I. Let β' be the rate at which the papers in SF are cited by members of N who are producing papers in SF. As the infection process develops,

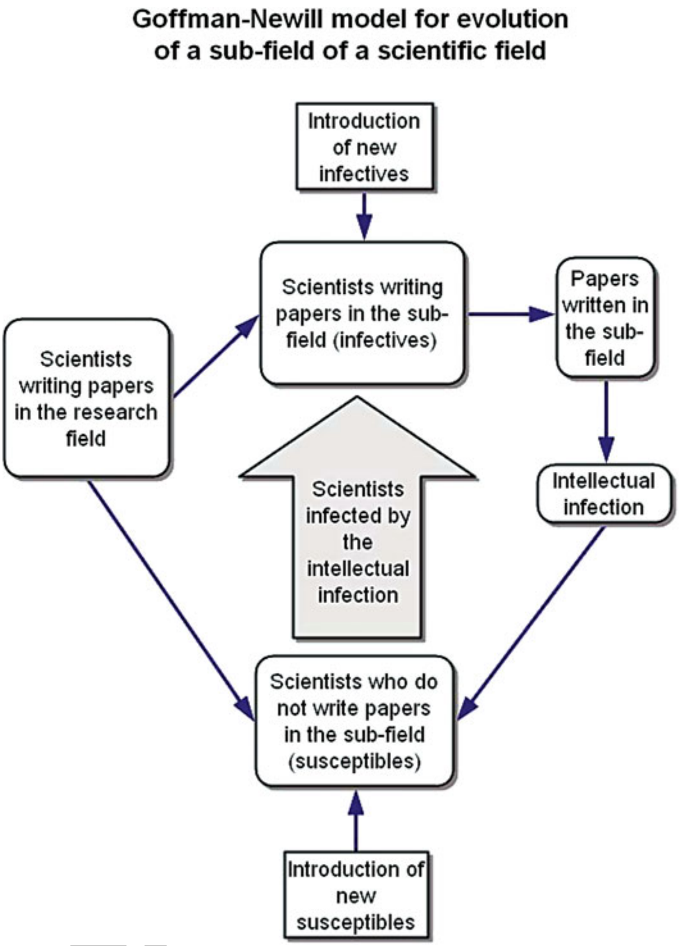


Fig. 3.14 Schema of Goffman–Newill model for the evolution of a scientific field. Scientists are attracted to a sub-field after being intellectually infected by papers from the sub-field

some susceptibles and infectives are removed, i.e. some scientists are no longer active, and some papers are not cited anymore. Let γ and γ' be the rates of removal of infectives from the populations I and I' respectively, and δ and δ' be the rates of removal from the populations of susceptibles S and S'. In addition, there can be a supply of infectives and susceptibles in N and N'. Let the rates of introduction of new susceptibles be μ and μ', i.e. the rates at which the new authors and new papers are introduced in F, and let the rates of introduction of new infectives be υ and υ', i.e. the rates at which new authors and new papers are introduced in SF. In addition, within a short time interval a susceptible can remain susceptible or can become an infective or be removed; the infective can remain an infective or can become a removal; and the removal remains a removed. The immunes remain immune and do not return to the population of susceptibles. If, in addition, the populations are homogeneously mixed, the system of model equations reads

$$\frac{dS}{dt} = -\beta S I' - \delta S + \mu; \qquad \frac{dI}{dt} = \beta S I' - \gamma I + \upsilon \tag{3.29}$$

$$\frac{dR}{dt} = \gamma I + \delta S; \qquad \frac{dS'}{dt} = -\beta' S' I - \delta S' + \mu' \tag{3.30}$$

$$\frac{dI'}{dt} = \beta' S' I - \gamma' I' + \upsilon'; \qquad \frac{dR'}{dt} = \gamma' I' + \delta' S'. \tag{3.31}$$

The conditions for development of an epidemic are as follows. If as an initial condition at t_0, a single infective is introduced into the populations N_0 and N'_0, then for an epidemic to develop, the change of the number of infectives must be positive in both populations. Then, for $\rho = \frac{\gamma - \upsilon}{\beta}$ and $\rho' = \frac{\gamma' - \upsilon'}{\beta'}$, the threshold for the epidemic arises from the conditions $\beta S I' > \gamma I - \upsilon$ and $\beta' S' I' > \gamma' I' - \upsilon'$, such that the threshold is

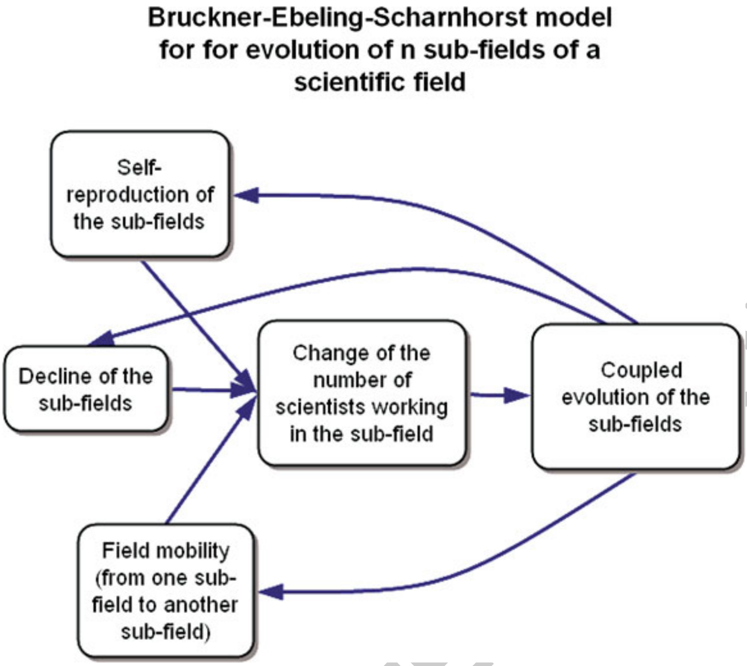
$$S_0 S'_0 > \rho \rho'. \tag{3.32}$$

The development of epidemics is given by the equation $\frac{dI}{dt} = D(t)$. The peaks of the epidemic occur at time points where $\frac{d^2 I}{dt^2} = 0$, while the epidemic's size is given by $I(t \to \infty)$.

(2) Bruckner–Ebeling–Scharnhorst model for the growth of *n* subfields in a scientific field

The evolution of growth processes in a system of scientific fields can be modeled by complex continuous evolution models. One of them, the Bruckner–Ebeling–Scharnhorst approach (Bruckner et al. 1990) (Fig. 3.15), is closely related to several generalizations of Eigen's theory of prebiotic evolution and is briefly discussed here (see also Ebeling et al. 2006). In 1912, Lotka (Lotka 1912) published the idea of describing biological epidemic processes, like malaria, as well as chemical oscillations, with the help of a set of differential equations. These equations, known as Lotka–Volterra equations (Lotka 1925; Volterra 1927), are used to describe a coupled growth process of populations. However, they do not reflect several essential properties of evolutionary processes such as the creation of new structural elements. Because of this, one has to consider a more general set of equations for the change in the number x_i of the scientists from the ith scientific subfield (a Fisher–Eigen–Schuster kind of model), i.e.,

$$\frac{dx_i}{dt} = (A_i - D_i)x_i + \sum_{j=1;j\neq i}^{n} (A_{ij}x_j - A_{ji}x_i) + \sum_{j=1;j\neq i}^{n} B_{ij}x_i x_j - k_0 x_i, \qquad i,j = 1,\ldots,n. \tag{3.33}$$



this figure will be printed in b/w

Fig. 3.15 Schema of Bruckner–Ebeling–Scharnhorst model of evolution of n scientific sub-fields. Self-reproduction and decline of subfields as well as field mobility are taken into account

The model based on (3.33) describes the coupled growth of n subfields, of a scientific discipline. Three fundamental processes of evolution are included in (3.33) : (a) self-reproduction: students and young scientists join the field and start working on corresponding problems. Their choice is influenced mainly by the education process as well as by individual interests and by existing scientific schools; (b) decline: scientists are active in science for a limited number of years. For different reasons (for example, retirement) they stop working and leave the system; (c) field mobility: individuals turn to other fields of research for various reasons or maybe open up new ones themselves.

The reasoning to obtain (3.33) goes as follows. The general form of the law for growth of the ith subfield is supposed to be

$$\frac{dx_i}{dt} = f_i(\mathbf{x}), \qquad \mathbf{x} = (x_1, \dots, x_n). \tag{3.34}$$

By separation, $f_i = w_i x_i$, one obtains the replicator equation

$$\frac{dx_i}{dt} = w_i x_i, \qquad i = 1, 2, \dots, n. \tag{3.35}$$

Notice that when $w_i = \text{const}$, the fields are uncoupled, i.e., there is an exponential growth in science. Otherwise, w_i itself is a function of x and of various parameters, but can be separated into three terms according to the above model assumptions, i.e.,

$$w_i = A_i - D_i + \sum_{j=1, j \neq i}^{n} \left(A_{ij} \frac{x_j}{x_i} - A_{ij} \right). \tag{3.36}$$

Equation (3.33) is thus obtained from (3.35) and (3.36) for $B_{ij} = 0, k_0 = 0$. To adapt this model to real growth processes, it can be assumed that the coefficients A_i, D_i, and A_{ij} themselves are functions of x_i:

$$A_i = A_i^0 + A_i^1 x_i + \ldots; \;\; D_i = D_i^0 + D_i^1 x_i + \ldots; \;\; A_{ij} = A_{ij}^0 + A_{ij}^1 x_j + \ldots \tag{3.37}$$

Each of the three fundamental processes of change is represented in (3.33) with a linear and a quadratic term only. For example, the terms A_i^1 and D_i^1 account for cooperative effects in self-reproduction and decline processes respectively, while D_i^0 accounts for a decline, because of aging. The contributions A_{ij}^0 assume a linear type of field mobility behavior for scientists analogous to a diffusion process. On the other hand, the terms A_{ij}^1 represent a directed process of exchange of scientists between fields. The best way to obtain these parameters is to estimate them for specific data bases using the method of least squares.

Key point Nr. 11
The Bruckner–Ebeling–Scharnhorst model does not belong to the class of epidemic models which are best applicable only for describing the expansion stage of a process. The Bruckner–Ebeling–Scharnhorst model is an evolution model: it describes all stages of the evolution of a system.

3.4 Small-Size Scientific and Technological Systems: Stochastic Models (Fig. 3.16)

The movement of large bodies in mechanics is governed by deterministic laws. When the body contains a small number of molecules and atoms, stochastic effects such as the Brownian motion become important. In the area of scientific systems, the fluctuations become very important when the number of scientists in a certain research subfield is small. This is typical for new research fields with only a few researching scientists.

Several examples of stochastic models for the description of the diffusion of ideas or technology and the evolution of science are: (a) the model of evolution of

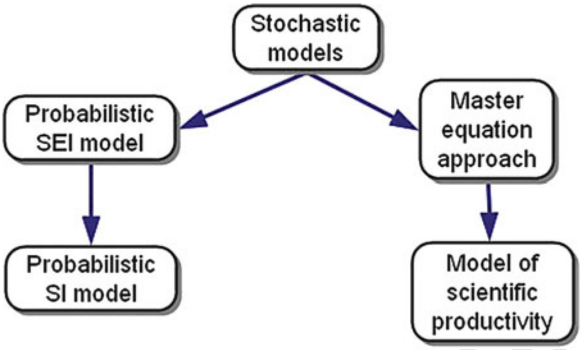


Fig. 3.16 Hierarchy of stochastic models discussed in this chapter

scientific disciplines with an example pertaining to the case of elementary particles physics (Kot 1987); (b) stochastic models for the aging of scientific literature (Glänzel and Schoepflin 1994); c) stochastic models of the Hirsch index (Burrell 2007) and of instabilities in evolutionary systems (Bruckner et al. 1989); (d) models of implementation of technological innovations (Bruckner et al. 1996), etc. (Braun et al. 1985). In the following, see Fig. 3.16, two probabilistic and two stochastic models are discussed. Some attention is devoted to the master equation approach as well.

3.4.1 *Probabilistic SI and SEI Models*

Epidemiological models of differential-equation-based compartmental type have been found to be limited in their capacity to capture heterogeneities at the individual level and in the interaction between individual epidemiological units (Chen and Hicks 2004). This is one of the reasons to switch from models in which the number of individuals are in given known states to models involving probabilities. One such model (Kiss et al. 2000) captures the diffusion of topics over a network of connections between scientific disciplines, as assigned by the ISI Web of Science's classification in terms of Subject Categories (SCs). Each SC is considered as a node of a network along with all its directed and weighted connections to other nodes or SCs (Kiss et al. 2000, 2005). As with epidemic models, nodes can be characterized in a medical way. SCs that are susceptible (S) are either not aware of a particular research topic or, if aware, may not be ready to adopt it. Incubating SCs (E) are those that are aware of a certain topic and have moved to do some research on problems connected with this topic. Infected SCs (I) are actively working and publishing in a particular research topic.

Two probabilistic models, i.e., (i) the Susceptible-Exposed-Infected (SEI) model (Fig. 3.17) and (ii) a simpler Susceptible-Infected (SI) model (Fig. 3.18), are thereby only discussed.

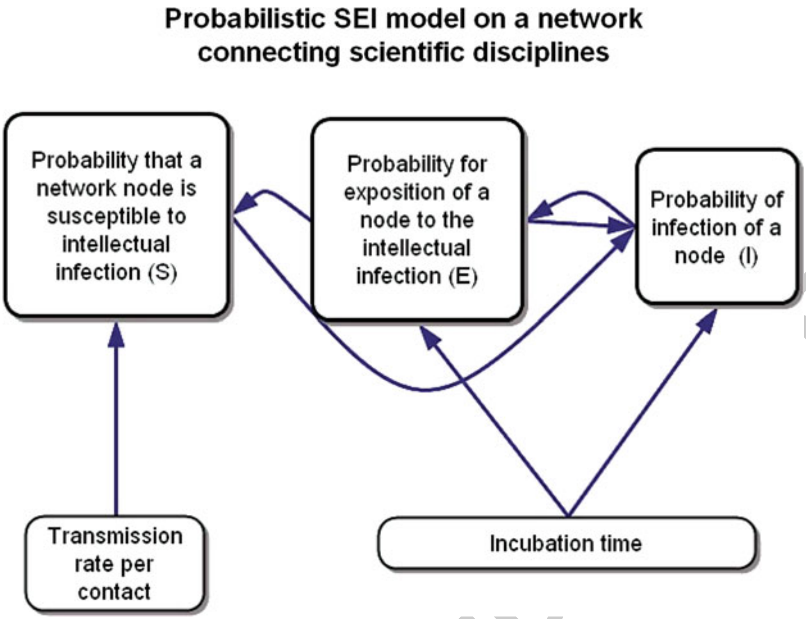
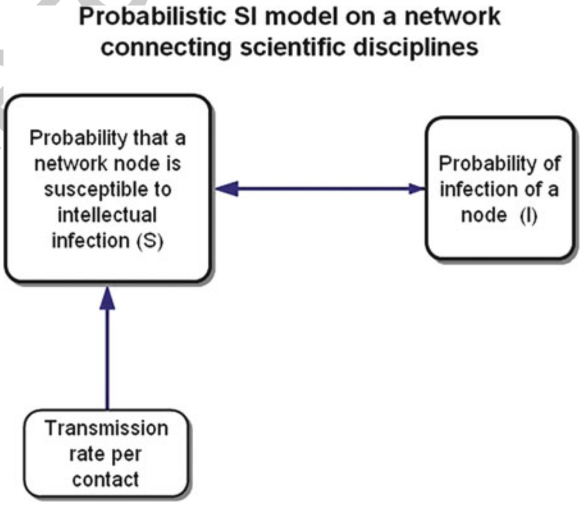


Fig. 3.17 Schema of the probabilistic SEI model for epidemics in a network connecting scientific disciplines

Fig. 3.18 Schema of the probabilistic SI model for epidemics in a network connecting scientific disciplines



(1) Susceptible-Exposed-Infected (SEI) model

The SEI model equations for the evolution of the node state probabilities are given by (Kiss et al. 2000):

$$\frac{dS_i(t)}{dt} = -\sum_j A_{ji} I_j(t) S_i(t), \tag{3.38}$$

$$\frac{dE_i(t)}{dt} = \sum_j A_{ji} I_j(t) S_i(t) - \gamma E_i(t), \tag{3.39}$$

$$\frac{dI_i(t)}{dt} = \gamma E_i(t), \tag{3.40}$$

where $0 \leq I_i(t) \leq 1$ denotes the probability of node i being infected at time t (likewise for $S_i(t)$ and $E_i(t)$). The directed and weighted contact network is represented by $A_{ij} = r\Gamma_{ij}$ with $\Gamma_{ij} = (w_{ij})_{i,j=1,\ldots,N}$ denoting the adjacency matrix that includes weighted links; r is the transmission rate per contact and $1/\gamma$ is the average incubation or latent period.

This set of equations states that an increase in the probability E_i of a node i being exposed to an infection is directly proportional to the probability S_i of node i being susceptible and the probability I_j of neighbouring nodes j being infected. The number of such contacts and the per-contact rate of transmission are incorporated in A_{ij}. Likewise, E_i decreases if exposed/infected nodes become infected after an average incubation time $1/\gamma$. The number of infected SCs at time t, according to the model, can be estimated as $I(t) = \sum_i I_i(t)$. Since $S_i(t) + E_i(t) + I_i(t) = 1$, for each $t > 0$, (3.38)–(3.40) are readily understood, in view of (3.39).

(2) Susceptible-Infected (SI) model

The above SEI model can be simplified to an SI model when the possibility of an exposed period is excluded, i.e,. if $\frac{dE_i(t)}{dt} = 0$. The equations for this simpler SEI model are reduced to

$$\frac{dS_i(t)}{dt} = -\sum_j A_{ji} I_j(t) S_i(t); \qquad \frac{dI_i(t)}{dt} = \sum_j A_{ji} I_j(t) S_i(t), \tag{3.41}$$

where the probability I_i of a node i being infected and infectious only depends on the probability S_i of the node i being susceptible. The comparison of both models with available data shows (Kiss et al. 2000) that while the agreement at the population level is usually much better for the SEI model, for the same pair of parameters, the agreement at the individual level is better when the simpler SI model is used.

3.4.2 Master Equation Approach

(1) Stochastic evolution model with self-reproduction, decline, and field mobility

There exists a high correlation between field mobility processes and the emergence of new fields (Bruckner et al. 1990). This can be accounted for by a stochastic model (see Fig. 3.19), in which the system at time t is characterized by a set of integers $N_1, N_2, \ldots, N_i, \ldots, N_n$, with N_i being, e.g., the number of scientists working in the subfield i, which is considered now as a stochastic variable. The three fundamental types of scientific change mentioned in the discussion of the Bruckner–Ebeling–Scharnhorst model (see above) here correspond to three elementary stochastic processes with three different transition probabilities:

(a) For self-reproduction, the transition probability is given by $W(N_i + 1 \mid N_i) = A_i^0 N_i = A_i^0 N_i + A_i^1 N_i (N_i - 1)$.

(b) The transition probability for decline is $W(N_i - 1 \mid N_i) = D_i^0 N_i + D_i^1 N_i (N_i - 1)$.

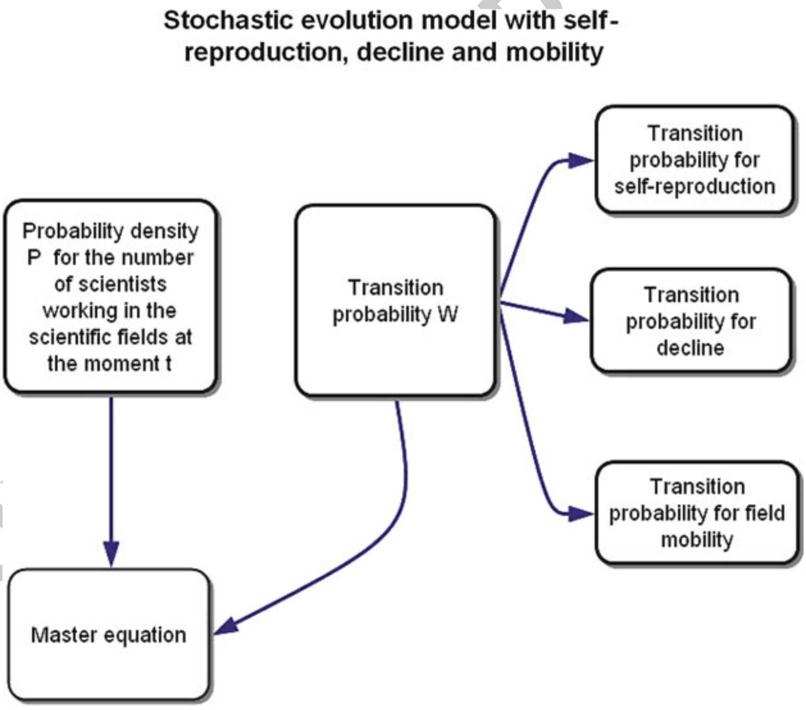


Fig. 3.19 Schema of the master equation model of evolution of scientific fields in presence of self-reproduction, decline, and field mobility

(c) The transition probability for field mobility is $W(N_i + 1, N_j - 1 \mid N_i N_j) = A^0_{ij} N_j + A^1_{ij} N_i N_j$.

The probability density $P(N_1, \ldots, N_i, N_j, \ldots, t)$ is given by the so-called master equation

$$\frac{\partial P}{\partial t} = WP \tag{3.42}$$

which can be solved analytically only in some very special cases (van Kampen 1981).

(2) The master equation as a model of scientific productivity
The productivity factor is a very important ingredient in mathematically simulating a scientific community evolution. One way to model such an evolution is through a dynamic equation which takes into account the stochastic fluctuations of scientific community members productivity (Romanov and Terekhov 1997)

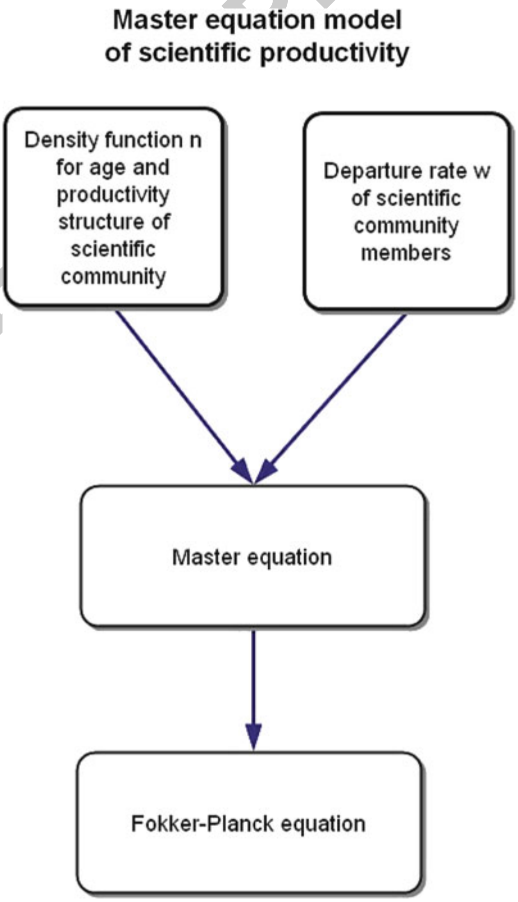


Fig. 3.20 Schema of the master equation model for scientific productivity

(Fig. 3.20). The main processes of scientific community evolution accounted for by this model are, beside the biological constraints (like the self-reproduction, aging of scientists, and death), their departure from the field due to mobility or abandon of research activities. Call a the age of an individual and let a scientific productivity index ξ be in incorporated into the individual state space; both a and ξ are being considered to be continuous variables with values in $[0, \infty]$. The scientific community dynamics is described by a number density function $n(a, \xi, t)$, – another form of scientific landscape, which specifies the age and productivity structure of the scientific community at time t. For example, the number of individuals with age in $[a_1, a_2]$ and scientific productivity in $[\xi_1, \xi_2]$ at time t is given by the integral $\int_{a_1}^{a_2} \int_{\xi_1}^{\xi_2} da \, d\xi \, n(a, \xi, t)$.

A master equation for this function $n(a, \xi, t)$ can be derived (Romanov and Terekhov 1997):

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n(a, \xi, t) = -[J(a, \xi, t) + w(a, \xi, t)]\, n(a, \xi, t) + \int_{-\infty}^{\xi} d\xi' \, \chi(a, \xi - \xi', t)\, n(a, \xi - \xi', t), \qquad (3.43)$$

where $w(a, \xi, t)$ denotes the departure rate of community members. If $x(t)$ is a random process describing the scientific productivity variation and if $p_a(x, t \mid y, \tau)$ (with $\tau < t$) is the transition probability density corresponding to such a process, then

$$\chi(a, \xi, \xi', t) = \lim_{\Delta t \to 0} \frac{p_a(\xi + \xi', t + \Delta t \mid \xi, t)}{\Delta t}. \qquad (3.44)$$

The transition rate, at time t from the productivity level ξ, $J(a, \xi, t)$ is by definition: $J(a, \xi, t) = \int_{-\xi}^{\infty} d\xi' \, \chi(a, \xi, \xi', t)$. The increment ξ' may be positive or negative. The balance equation for $n(a, \xi, t)$ reads as follows

$$n(a + \Delta a, \xi, t + \Delta t) = n(a, \xi, t) - J(a, \xi, t)\, n(a, \xi, t)\, \Delta t + \left[\int_{-\infty}^{\xi} \chi(a, \xi - \xi', t)\, n(a, \xi - \xi', t) d\xi'\right] \Delta t - w(a, \xi, t)\, n(a, \xi, t)\, \Delta t. \quad (3.45)$$

The term on the right-hand side, $[1 - J(a, \xi, t)\Delta t] n(a, \xi, t)$, describes the proportion of individuals whose scientific productivity does not change in $[t, t + \Delta t]$; the integral term describes the individuals whose scientific productivity becomes equal to ξ because of increasing or decreasing in $[t, t + \Delta t]$; the last term corresponds to the departure of individuals due to stopping research activities or death. After expanding $n(a + \Delta t, \xi, t + \Delta t)$ around a and t, keeping terms up to the first order in Δt, one obtains the master equation (3.43).

As the master equation is difficult to handle for an elaborate analysis, it is often reduced to an approximated equation similar to the well-known Fokker–Planck equation (Risken 1984; Hänggi and Thomas 1982; Gardiner 1983). The approximation goes as follows. Let

$$\mu_k(a,\xi,t) = \int_{-\xi}^{\infty} d\xi'(\xi')^k \chi(a,\xi,\xi',t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} < (\xi')^k >; \; k = 1,2,\ldots, \tag{3.46}$$

where the brackets denote the average with respect to the conditional probability density $p_a(\xi + \xi', t + \Delta t \mid \xi, t)$. In addition, the following assumptions are made: (i) $\mu_1, \mu_2 < \infty$; $\mu_k = 0$ for $k > 3$; (ii) $n(a,\xi,t)$ and $\chi(a,\xi,\xi',t)$ are analytic in ξ for all a, t and ξ'. The additional assumption $\mu_k = 0$ for $k > 3$ demands the productivity to be continuous in the sense that as $\Delta t \to 0$, the probability of large fluctuations $\mid \xi' \mid$ must decrease so quickly that $<\mid \xi' \mid^3> \to 0$ more quickly than Δt.

When the above assumptions hold, the function n satisfies the equation (Romanov and Terekhov 1997):

$$\left(\frac{\partial}{\partial a} + \frac{\partial}{\partial t}\right) n = -\frac{\partial(\mu_1 n)}{\partial \xi} + \frac{1}{2}\frac{\partial^2(\mu_2 n)}{\partial \xi^2} - wn. \tag{3.47}$$

If $w = 0$, (3.47) is converted to the well known Fokker–Planck equation. (3.47) describes the scientific community evolution through a drift along the age component and a drift and diffusion with respect to the productivity component. The diffusion term characterized by the diffusivity μ_2 takes into account the stochastic fluctuations of scientific productivity conditioned by internal factors (such as individual abilities, labour motivations, etc.) and external factors (such as labor organization, stimulation system, etc.). The initial and boundary conditions for (3.47) are: (a) $n(a,\xi,0) = n^0(a,\xi)$, where $n^0(a,\xi)$ is a known function defining the community age and productivity distribution at time $t = 0$; and (b) $n(0,\xi,t) = \nu(\xi,t)$ where the function $\nu(\xi,t)$ represents the intensity of input flow of new members at age $a = 0$ being set $\nu(\xi,0) = n^0(0,\xi)$. In addition, $n(a,\xi,t) \to 0$ as $a \to \infty$.

The general solution of equation (3.47) with the above initial condition (a) and boundary condition (b) is still a difficult task. However, for many practical applications, a knowledge of first and second moments of distribution function $n(a,\xi,t)$ is sufficient. Equation (3.47) can be solved numerically or can be reduced to a system of ordinary differential equations (Romanov and Terekhov 1997).

Finally, two additional problems that can be treated by the master equation approach can be mentioned:

- Age-dependent models where the birth and death rates connected to the selection are age-dependent (Ebeling et al. 1986, 1990)
- The problem of new species in evolving networks (Ebeling et al. 2006). On the basis of a stochastic treatment of the problem, the notion of 'innovation' can be introduced in a broad sense as a disturbance and/or an instability of a corresponding social, technological, or scientific system. The fate of a small number of individuals of a new species in a biological system can be thought to be mathematically equivalent to some extent to the fate of a new idea, a new technology, or a new model of behavior. The evolution of the new species can be studied on evolving networks, where some nodes can disappear and new

nodes can be introduced. This evolution of the network can change significantly the dynamic behavior of the entire system of interacting species itself. Some of the species can vanish in a finite time. This feature can be captured effectively by the master equation approach.

Key point Nr. 12
In deterministic cases, the system is robust against fluctuations: it follows some trajectory and the fluctuations are too weak to change it. When the fluctuations are important, then different trajectories for the evolution of the system become possible. To each trajectory, a probability can be assigned. This probability reflects the chance that the system will follow the corresponding trajectory. The collection of the probabilities leads to a probability distribution which can be calculated, in many evolutionary cases, on the basis of the master equation approach.

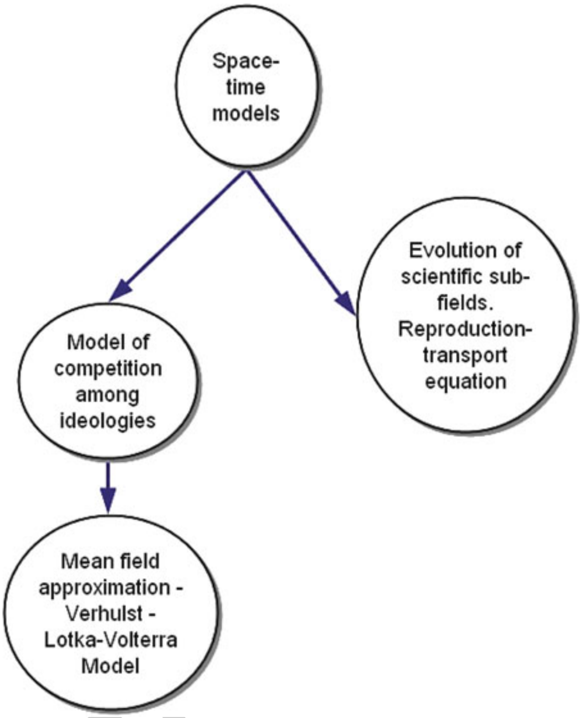
3.5 Space-Time Models: Competition of Ideas – Ideological Struggle

A further level of complication is to include spatial variables explicitly in the above models describing the diffusion of ideas. At this stage of globalization of economies, with several of its concomitant features, like idea, knowledge, and technology diffusion, to consider the spatial aspect is clearly a must. A large amount of research on the spatial aspects of diffusion of populations is already available. As examples of early work, papers by Kerner (1959); Allen (1975); Okubo (1980), and Willson and de Roos (1993) can be pointed out. From the point of view of diffusion of ideas and scientists, the previously discussed continuous model of research mobility (Bruckner et al. 1990) has to be singled out. Moreover, the model presented below is closely connected to the space-time models of migration of populations developed by Vitanov and co-authors (Vitanov et al. 2009a,b). In addition, a reproduction-transport equation model (see Fig. 3.21) can be discussed.

3.5.1 Model of Competition Between Ideologies

The diffusion of ideas is necessarily accompanied by competition processes. One model of competition between systems of ideas (ideologies) goes as follows (Fig. 3.22). Let a population of N individuals occupy a two-dimensional plane. Suppose that there exists a set of ideas or ideologies $P = \{P_0, P_1, \dots, P_n\}$ and let N_i members of the population be followers of the P_i ideology. The members N_0

Fig. 3.21 Relations between space-time models discussed in this chapter



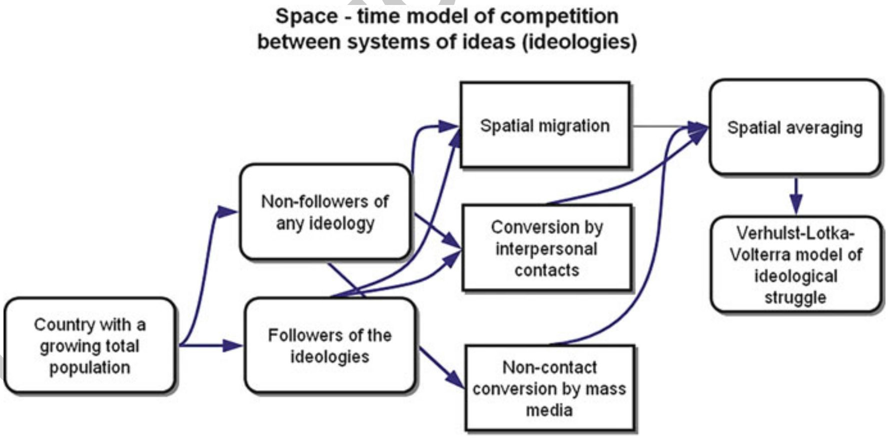


Fig. 3.22 Schema of the space-time model for describing competition between ideologies

of the class P_0 are not supporters of any ideology; in some sense, they have their own individual one and do not wish to be considered associated with another one, global or not. In such a way, the population is divided in $n + 1$ sub-populations of followers of different ideologies. The total population is: $N = N_0 + N_1 + \ldots N_n$.

Let a small region $\Delta S = \Delta x \Delta y$ be selected in the plane. In this region there are ΔN_i individuals holding the ith ideology, $i = 0, 1, \ldots, n$. If ΔS is sufficiently small, the density of the ith population can be defined as $\rho_i(x, y, t) = \frac{\Delta N_i}{\Delta S}$.

Allow the members of the ith population to move through the borders of the area ΔS. Let $\mathbf{j}_i(x, y, t)$ be the current of this movement. Then $(\mathbf{j}_i \cdot \mathbf{n})\delta l$ is the net number of members of the ith population/ideology, crossing a small line δl with normal vector $\mathbf{n}$. Let the changes be summarized by the function $C_i(x, y, t)$. The total change of the number of members of the ith population is

$$\frac{\partial \rho_i}{\partial t} + \mathrm{div}\mathbf{j}_i = C_i. \tag{3.48}$$

The first term in (3.48) describes the net rate of increase of the density of the ith population. The second term describes the net rate of immigration into the area. The r.h.s. of (3.48) describes the net rate of increase exclusive of immigration.

Let us now specify $\mathbf{j}_i$ and C_i: $\mathbf{j}_i$ is assumed to be made of a non-diffusion part $\mathbf{j}_i^{(1)}$ and a diffusion part $\mathbf{j}_i^{(2)}$ where $\mathbf{j}_i^{(2)}$ is assumed to have the general form of a linear multicomponent diffusion (Kerner 1959) in terms of a diffusion coefficient D_{ik}

$$\mathbf{j}_i = \mathbf{j}_I^{(1)} + \mathbf{j}_2^{(2)} = \mathbf{j}_i^{(1)} - \sum_{k=0}^{n} D_{ik}(\rho_i, \rho_k, x, y, t)\nabla \rho_k. \tag{3.49}$$

Let some of the followers of the ideology P_i be capable of and interested in changing ideology: i.e., they can convert from the ideology P_i to the ideology P_j. It can be assumed that the following processes can happen with respect to the members of the subpopulations of the property holders: (a) deaths: described by a term $r_i \rho_i$. It is assumed that the number of deaths in the ith population is proportional to its population density. In general $r_i = r_i(\rho_\nu, x, y, t; p_\mu)$, where ρ_ν stands for $(\rho_0, \rho_1, \ldots, \rho_N)$ and p_μ stands for $(p_1, \ldots, p_M)$ containing parameters of the environment; (b) non-contact conversion: in this class are included all conversions exclusive of the conversions by interpersonal contact between the members of whatever populations. A reason for non-contact conversion can be the existence of different kinds of mass communication media which make propaganda for whatever ideologies. As a result, members of each population can change ideology. For the ith population, the change in the number of members is: $\sum_{j=0}^{n} f_{ij}\rho_j$, $f_{ii} = 0$. In general, $f_{ij} = f_{ij}(\rho_\nu, x, y, t; p_\mu)$; (c) contact conversion: it is assumed that there can be interpersonal contacts among the population members. The contacts happen between members in groups consisting of two members (binary contacts), three members (ternary contacts), four members, etc. As a result of the contacts, members of each population can change their ideology. For binary contacts, let it be assumed that the change of ideology probability for a member of the jth population is proportional to the possible number of contacts, i.e., to the density of the ith population. Then the total number of "conversions" from P_j to P_i is $a_{ij}\rho_i\rho_j$, where a_{ij} is a parameter. In order to have a ternary contact, one must have a group of three members. The most simple is to assume that such a group exists with a probability proportional to the corresponding densities of the concerned populations. In a ternary contact between members of the ith, jth, and kth population, members of the jth and kth populations can change their ideology according to $P_i = b_{ijk}\rho_i\rho_j\rho_k$, where b_{ijk} is a parameter. In general, $a_{ij} = a_{ij}(\rho_\nu, x, y, t; p_\mu)$; $b_{ijk} = b_{ijk}(\rho_\nu, x, y, t; p_\mu)$; etc.

On the basis of the above, the C_i term looks as follows (for more research of these types of population models see (Dimitrova and Vitanov 2000, 2001a,b)):

$$C_i = r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j + \sum_{j=0}^{n} a_{ij}\rho_i\rho_j + \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \dots, \qquad (3.50)$$

and the model system of equations becomes

$$\begin{aligned}\frac{\partial\rho_i}{\partial t} + \mathrm{div}\mathbf{j}_i^{(1)} - \sum_{j=0}^{n}\mathrm{div}(D_{ij}\nabla\rho_j) &= r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j \\ &+ \sum_{j=0}^{n} a_{ij}\rho_i\rho_j + \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \dots \qquad (3.51)\end{aligned}$$

The density of the entire population is $\rho = \sum_{i=0}^{n}\rho_i$. It can be assumed that it changes in time according to the Verhulst law (but see the note after (3.56)!)

$$\frac{\partial\rho}{\partial t} = r\rho\left(1 - \frac{\rho}{C}\right) \qquad (3.52)$$

where $C(\rho_\nu, x, y, t; p_\mu)$ is the so-called carrying capacity of the environment (Odum 1959) and $r(\rho_\nu, x, y, t; p_\mu)$ is a positive or negative growth rate. When pertinent sociological data are available, the same type of equation could hold for any ith population with a given r_i.

First, consider the case in which the current $\mathbf{j}_i^{(i)}$ is negligible, i.e., $\mathbf{j}_i^{(i)} \approx 0$ (*no diffusion approximation*). In addition, consider only the case when all parameters are constants. The model system of equations becomes

$$\begin{aligned}\frac{\partial\rho_i}{\partial t} - D_{ij}\sum_{j=0}^{n}\Delta\rho_j &= r_i\rho_i + \sum_{j=0}^{n} f_{ij}\rho_j + \sum_{j=0}^{n} a_{ij}\rho_i\rho_j \\ &+ \sum_{j,k=0}^{n} b_{ijk}\rho_i\rho_j\rho_k + \dots, \qquad (3.53)\end{aligned}$$

for

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \qquad i = 0, 1, 2, \dots, n. \qquad (3.54)$$

Let plane-averaged quantities and fluctuations (linear or nonlinear) be enough relevant. Let $q(x, y, t)$ be a quantity defined in an area S. By definition, a plane-averaged quantity is $\overline{q} = \frac{1}{S}\int\int_S dxdy\ q(x, y, t)$. Call the fluctuations $Q(x, y, t)$ such that $q(x, y, t) = \overline{q}(t) + Q(x, y, t)$. If the territory is large and within the stationary approximation, S can be assumed to be large enough such that each plane-averaged combination of fluctuations vanishes, such that $\overline{Q_i} = \overline{Q_i Q_j} = \overline{Q_i Q_j Q_k} = \cdots = 0$. In addition to S being large and $\int\int_S dxdy \Delta Q_k$ assumed to be finite, it can be also assumed that $\overline{\Delta Q_k} = \frac{1}{S}\int\int_S dxdy\Delta Q_k \to 0$.

On the basis of the above (reasonable) assumptions, it is possible to separate the dynamics of the averaged quantities from the dynamics of fluctuations. As a result of the plane-average of (3.53), the following equations for the dynamics of the plane-averaged densities are obtained

$$\overline{\rho}_0 = \overline{\rho} - \sum_{i=1}^{n} \overline{\rho}_i; \quad \frac{d\overline{\rho}}{dt} = r\overline{\rho}\left(1 - \frac{\overline{\rho}}{C}\right) \tag{3.55}$$

$$\frac{d\overline{\rho}_i}{dt} = r_i\overline{\rho}_i + \sum_{j=0}^{n} f_{ij}\overline{\rho}_j + \sum_{j=0}^{n} a_{ij}\overline{\rho}_i\overline{\rho}_j + \sum_{j,k=0}^{n} b_{ijk}\overline{\rho}_i\overline{\rho}_j\overline{\rho}_k + \ldots \tag{3.56}$$

Instead of (3.55) we can write an equation for $\overline{\rho}_0$ from the kind of (3.56). Then the total population density $\overline{\rho}$ will not follow the Verhulst law.

Equations (3.55) and (3.56) represent the model of ideological struggle proposed by Vitanov et al. (2010). There is one important difference between the Lotka–Volterra models (Lotka 1912; Volterra 1927), often used for describing prey-predator systems, and the above model of ideological struggle. The originality resides in the generalization of usual prey-predator models to the case in which a prey (or predator) changes its state and becomes a member of the predator pack (or prey band), due to some interaction with its environment or with some other prey or predator. Indeed, it can be hard for rabbits and foxes to do so, but it can be often the case in a society: a member of one population can drop his/her ideology and can convert to another one.

In order to show the relevance of such extra conditions on an evolution of populations, consider a huge (mathematical) approximation – it might be a drastic one in particular in a country with a strictly growing total population. (Recall that the growth rate r could be positive or negative or time-dependent). Let r be > 0 and let the maximum possible population of the country be C. Consider more convenient notations by setting $\overline{\rho} = N$; $\overline{\rho}_0 = N_0$; $\overline{\rho}_i = N_i$ and assume that the binary contact conversion is much stronger than the ternary, etc. conversions. The system equations become

$$N = N_0 + \sum_{i=1}^{n} N_i; \qquad \frac{dN}{dt} = rN\left(1 - \frac{N}{C}\right) \tag{3.57}$$

$$\frac{dN_i}{dt} = r_i N_i + \sum_{j=0}^{n} f_{ij} N_j + \sum_{j=0}^{n} b_{ij} N_i N_j. \tag{3.58}$$

Reduce the discussion of (3.57) and (3.58) to a society in which there is the spreading of only one ideology; therefore, the population of the country is divided into two groups: N_1, followers of the “invading” ideology and N_0, people who are at first “indifferent” to this ideology. Let only the non-contact conversion scheme exist, as possibly moving the ideology-free population toward the single ideology; thus f_{10} is finite, but $b_{10} = 0$. Let the initial conditions be $N(t = 0) = N(0)$ and $N_1(t = 0) = N_1(0)$. The solution of the system of model equations is

$$N(t) = \frac{CN(0)}{N(0) + (C - N(0))e^{-rt}}, \tag{3.59}$$

like the Verhulst law, but

$$N_1(t) = e^{-(f_{10}-r_1)t}\left\{N_1(0) + \frac{Cf_{10}}{r}\left[\Phi\left(-\frac{C-N(0)}{N(0)}, 1, -\frac{f_{10}-r_1}{r}\right)\right.\right.$$
$$\left.\left. -e^{(f_{10}-r_1)t}\Phi\left(-\frac{C-N(0)}{N(0)e^{rt}}, 1, -\frac{f_{10}-r_1}{r}\right)\right]\right\} \tag{3.60}$$

with

$$N_0(t) = N(t) - N_1(t) \tag{3.61}$$

in which Φ is the special function $\Phi(z, a, v) = \sum_{n=0}^{\infty} \frac{z^n}{(v+n)^a}; \quad |z| < 1$. The obtained solution describes an evolution in which the total population N reaches asymptotically the carrying capacity C of the environment. The number of adepts of the ideology reaches an equilibrium value which corresponds to the fixed point $\hat{N}_1 = Cf_{10}/(f_{10} - r_1)$ of the model equation for $\frac{dN_1}{dt}$. The number of people who are not followers of the ideology asymptotically tends to $N_0 = C - \hat{N}_1$. Let $C = 1$, $f_{10} = 0.03$, and $r_1 = -0.02$, then $\hat{N}_1 = 0.6$, which means that the evolution of the system leads to an asymptotic state in which 60% of the population are followers of the ideology and 40% are not.

Other more complex cases with several competing ideologies can be discussed, observing steady states or/and cycles (with different values of the time intervals for each growth or/and decay), chaotic behaviors, etc. (Vitanov et al. 2010). In particular, it can be shown that accepting a slight change in the conditions of the environment can prevent the extinction of some ideology. After almost collapsing, some ideology can spread again and can affect a significant part of the country's population. Two kinds of such resurrection effects have been found and described as *phoenix effects* in the case of two competing ideologies. In the phoenix effect of the first kind, the equilibrium state connected to the extinction of the second ideology exists but is unstable. In the phoenix effect of the so-called second kind, the equilibrium state connected to extinction of the second ideology vanishes. *In fine*, the above model seems powerful enough to discuss many realistic cases. The number of control parameters seems huge, but that is the case for many competing epidemics in complex systems. However, it was observed that the values of parameters can be monitored when enough data is available, including the time scales (Vitanov et al. 2010).

Key point Nr. 13
Space-time models are very appropriate for modeling migration processes such as the spatial migration of scientists, besides the diffusion of ideas through competition without strictly physical motion.

3.5.2 Continuous Model of Evolution of Scientific Subfields: Reproduction-Transport Equation

The change of subject of a scientist can be considered as a migration process (Bruckner et al. 1990; Ebeling and Scharnhorst 2000). Let research problems be represented by sequences of signal words or macro-terms $P_i = (m_i^1, m_i^2, \dots, m_i^k, \dots, m_i^n)$ which are registered according to the frequency of their appearance, joint appearance, etc., respectively, in the texts. Each point of the problem space, described by a vector $\mathbf{q}$, corresponds to a research problem, with the problem space consisting of all scientific problems (no matter whether they are under investigation or not). The scientists distribute themselves over the space of scientific problems with density $x(\mathbf{q}, t)$. Thus, there is a number $x(\mathbf{q}, t) d\mathbf{q}$ working at time t in the element $d\mathbf{q}$. The field mobility processes correspond to a density change of scientists in the problem space: instead of working on problem $\mathbf{q}$, a scientist may begin to work on problem $\mathbf{q}'$. As a result, $x(\mathbf{q}, t)$ decreases and $x(\mathbf{q}', t)$ increases. This movement of scientists (see also Fig. 3.23) can be described by means of the following reproduction-transport-equation:

$$\frac{\partial x(\mathbf{q},t)}{\partial t} = x(\mathbf{q},t)\, w(\mathbf{q} \mid x) + \frac{\partial}{\partial \mathbf{q}} \left(f(\mathbf{q},x) + D(\mathbf{q}) \frac{\partial x(\mathbf{q},t)}{\partial \mathbf{q}} \right). \tag{3.62}$$

In (3.62), self-reproduction and decline are represented by the term $w(\mathbf{q} \mid x)\, x(\mathbf{q}, t)$. For the reproduction rate function $w(\mathbf{q} \mid x)$, one can write

$$w(\mathbf{q} \mid x) = a(\mathbf{q}) + \int d\mathbf{q}'\, b(\mathbf{q}, \mathbf{q}')\, x(\mathbf{q}', t). \tag{3.63}$$

The local value of $a(\mathbf{q})$ is an expression of the rate at which the number of scientists on field $\mathbf{q}$ is modified through self-reproduction and/or decline, while $b(\mathbf{q}, \mathbf{q}')$ describes the influence exerted on the field $\mathbf{q}$ by the neighbouring field $\mathbf{q}'$. The field

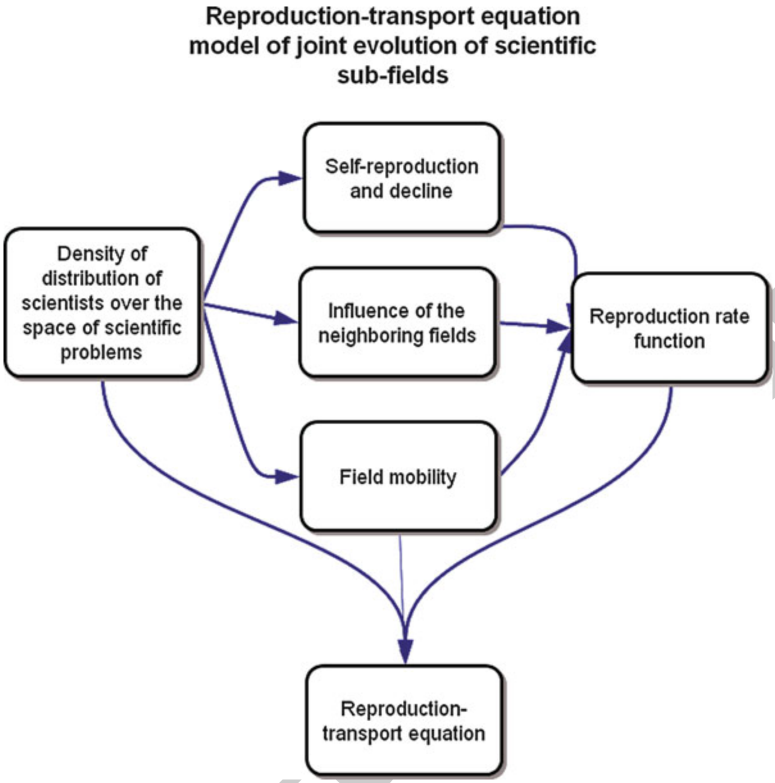


Fig. 3.23 Schema of the reproduction-transport equation model of joint evolution of scientific fields

mobility is modeled by means of the term $\frac{\partial}{\partial \mathbf{q}}\left(f(\mathbf{q},x)+D(\mathbf{q})\frac{\partial}{\partial \mathbf{q}}x(\mathbf{q},t)\right)$. In most cases, (3.62) can only be solved numerically. For more details on the model, see Bruckner et al. (1990).

3.6 Statistical Approaches to the Diffusion of Knowledge

Solomon and Richmond (2001, 2002) have shown that the systems of generalized Lotka–Volterra equations are closely connected to the Pareto–Zipf probability distribution. Since such a distribution arises among other distributions and laws connected to the description of the diffusion of knowledge, it is of interest to discuss briefly the diffusion of knowledge within statistical approach studies. Lotka was its pioneer; a large amount of research has followed. Just as examples, one can mention the work of Yablonsky and Haitun on the Lotka law for the distribution of scientific

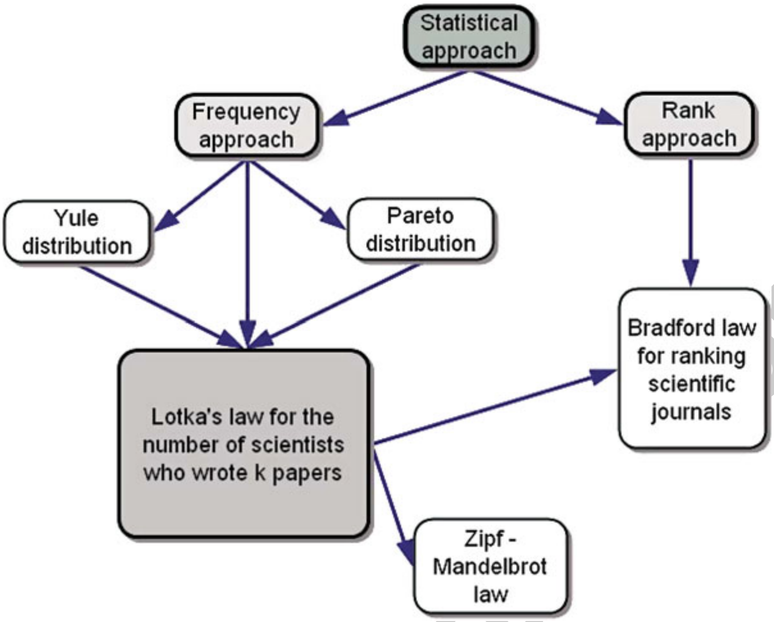


Fig. 3.24 Statistical laws and their relationships as discussed in the chapter

productivity and its connection with the Yule distribution (Yablonsky 1980, 1985; Haitun 1982), where the non-Gaussian nature of the scientific activities is emphasized. Interesting applications of the Zipf law are also presented in (Li 2002). The connection to the non-Gaussian distributions concepts of self-similarity and fractuality have been applied to the scientific system in (Katz 1999) and (van Raan 2000). Several tools for appropriate statistical analysis are hereby discussed. At the center of the discussion Lotka law shall receive some special attention (see Fig. 3.24).[3]

As part of this discussion on the statistical approach, the analysis of the productivity of scientists can be considered. The information connected to new ideas is thought to be often codified in scientific papers. Thus, the statistical aspects of scientific productivity is of practical importance. For example, the Lotka law reflects the distribution of publications over the set of authors considered as the information sources. Bradford law describes the distribution of papers on a given topic over the set of journals publishing these papers and ranked according to the order in the decrease of the number of papers on a given topic in each journal. These laws have a non-Gaussian nature and, because of this, possess specific features such as a concentration and dispersal effect (Yablonsky 1980): for example, it is found

---

[3] Let us mention a curious and interesting fact connected to statistical indicators. Very interesting is the conclusion in Gao and Guan (2009) that the scale-independent indicators show that in the fast growing innovation system of China, research institutions financed by the government play a more important role than the enterprises.

that there is a small number of highly productive scientists who write most of the papers on a given topic and, on the other hand, a large number of scientists with low productivity.

In order to give an example of the connection between the deterministic and statistical approaches, remember that the Goffman–Newill model, discussed here above, presents a connection between the number of scientists working in a research area and the number of relevant publications. In Bettencourt et al. (2008), it was found that the number of new publications scale as a simple power law with the corresponding number of new authors: $\Delta P = C(\Delta T)^{\alpha}$ where ΔP and ΔT are the new publications and the new authors over some time period (for an example 1 year). C is a normalization constant, and α is a scaling exponent. It has been demonstrated (Bettencourt et al. 2008) that the latter relationship provides a very good fit to data for six different research fields, but with different values of the scaling exponent α. For $\alpha > 1$, a field would grow by showing an increase in the number of publications per capita, i.e., in such a research field, the individual productivity increases as the field attracts new scientists. A field with $\alpha < 1$ has a per capita decrease in productivity. This can be a warning signal for a dying subject matter. It would be interesting to observe whether the exponent α is time-dependent, as is the case in related characterizing scaling exponents of financial markets (Vandewalle and Ausloos 1997) or in meteorology (Ivanova and Ausloos 1999). Policy control can thus be implemented for shaking α, thus the field mobility.

Key point Nr. 14
There exist two different kinds of statistical approaches for the analysis of scientific productivity: (i) the frequency approach and (ii) the rank approach. The frequency approach is based on the direct statistical counting of the number of corresponding information sources, such as scientists or journals. The rank approach is based on a ranking of the sources with respect to their productivity. The frequency and the rank approaches represent different and complementary reflections of the same law and form.

3.6.1 Lotka Law: Distributions of Pareto and Yule

Pareto (Chen et al. 1993) formulated the 80/20 rule: it can be expected that 20% of people will have 80% of the wealth. Or it can be expected that 80% of the citations refer to a core of 20% of the titles in journals. The idea of the rule of Pareto is very close to the research of Lotka who noticed the following dependence for the number of scientists n_k who wrote k papers

$$n_k = \frac{n_1}{k^2}; \quad k = 1, 2, \ldots, k_{max}. \tag{3.64}$$

In (3.64), n_1 is the number of scientists who wrote just one paper and k_{max} is the maximal productivity of a scientist.

$$\sum_{k=1}^{k_{max}} n_k = n_1 \sum_{k=1}^{k_{max}} \frac{1}{k^2} = N \tag{3.65}$$

where N is the total number of scientists. If we assume that $k_{max} \to \infty$ and take into account the fact that $\sum_{k=1}^{\infty} 1/k^2 = \pi/6$, we obtain a limiting value for the portion of scientists with the minimal productivity (single paper authors) in the given population of authors: $P_1 = n_1/N \approx 0.6$. Then, if the left and the right hand sides of (3.64) are divided by N, the frequency expression for the productivity distribution is: $p_1 = 0.6/k^2$; $\sum_{k=1}^{\infty} p_k = 1$. Equation (3.64) is called Lotka law, or the law of inverse squares: the number of scientists who wrote a given number of papers is inversely proportional to the square of this number of papers.

It must be noted that, like many other statistical regularities, Lotka law is valid only on the average since the exponent in the denominator of (3.64) is not necessarily equal to two (Yablonsky 1980). Thus, Lotka law should be considered as the most typical among a more general family of distributions:

$$n_k = \frac{n_1}{k^{1+\alpha}}; \qquad p_1 = \frac{p_1}{k^{1+\alpha}} \tag{3.66}$$

where α is the characteristic exponent of the distribution, n_1 is the normalizing coefficient which is determined as follows:

$$p_1 = \frac{n_1}{N} = \left(\sum_{k=1}^{k_{max}} \frac{1}{1+k^{\alpha}} \right)^{-1} . \tag{3.67}$$

Then the distribution of scientific output, (3.66), is determined by three parameters: the proportion of scientists with the minimal productivity p_1, the maximal productivity of a scientist k_{max}, and the characteristic exponent α. If one of these parameters is fixed, it is possible to study the dependence between two others. Let us fix k_{max} in (3.67). Then, we obtain the proportion of "single paper authors" p_1 as a function of α: $p_1(\alpha)$. When (3.67) is differentiated with respect to α, one can show that the corresponding derivative is positive for any $\alpha : dp_1(\alpha)/d\alpha > 0$. On the basis of a similar analysis of the portion of scientists with a larger productivity $p_k(\alpha)$ as a function of α, we arrive at the conclusion: the increase of α is accompanied by the increase of low-productivity scientists. This means that when the total number of scientists is preserved the portion of highly productive scientists will decrease.

Let us show that the Lotka law is an asymptotic expression for the Yule distribution. In order to obtain the Yule distribution, one considers the process of formation of a collection of publications as a Markov-type stochastic process. In addition, it is assumed that the probability of writing a new paper depends on the number of papers that have been already written by the scientist at time t: the probability of the transition into a new state on the interval $[t, t + \Delta t]$ should be a function of the state in which the system is at time t. Moreover, the probability of publishing a new paper during a time interval Δt, $p(x \to x + 1, \Delta t)$ is assumed to be proportional to the number x of papers that have been written by the scientists, introducing an intensity coefficient λ: $p(x \to x + 1, \Delta t) \propto \lambda x \Delta t$. After solving the corresponding system of differential equations for this process, the following expression (the Yule distribution) for the probability $p(x/t)$ of a scientist writing x papers during a time t is obtained (Yablonsky 1980):

$$p(x/t) = \exp(-\lambda t)(1 - \exp(-\lambda t))^{x-1}, \quad x = 1, 2 \ldots \tag{3.68}$$

The mean value of the Yule distribution is $x_t = \exp(\lambda t)$. Let us take into account the fact that every scientist works on a given subject during a certain finite random time interval $[0, t]$ which depends on the scientist's creative potential, the conditions for work, etc. With the simplest assumption that the probability of discontinuing work on a given subject is constant at any time, one obtains an exponential distribution for the time of work of any author in the scientific field under study: $p(t) = \mu \exp(-\mu t)$, where μ is the distribution parameter. The time parameter t which characterizes the productivity distribution, (3.68), is a random number. Then in order to obtain the final distribution of scientific output observed in the experiment over sufficiently large time intervals, (3.68) should be averaged with respect to this parameter t which is distributed according to the exponential law:

$$p(x) = \int_0^\infty dt\ p(x/t)p(t) = \int_0^\infty dt\ \exp(-\lambda t)(1 - \exp(-\lambda t))\mu \exp(-\mu t). \tag{3.69}$$

After integrating (3.69), the distribution of scientific output reads

$$p(x) = \frac{\mu}{\lambda} B\left(x, \frac{\mu}{\lambda} + 1\right) = \alpha B(x, \alpha + 1), \quad x = 1, 2, \ldots \tag{3.70}$$

where $B(x, \alpha + 1) = \Gamma(x)\Gamma(\alpha x + 1)/\Gamma(x + \alpha + 1)$ is a Beta-function, $\Gamma(x) \approx (x - 1)!$ is a Gamma-function, and $\alpha = \mu/\lambda$ is the characteristic exponent. For instance, if $\alpha \approx 1$ then $p(x) = 1/[x(x + 1)]$. Let us assume that $x \to \infty$ and apply the Stirling formula. Thus, the asymptotics of the Yule distribution (3.70) is like Lotka law (3.66) (up to a normalizing constant): $p(x) \propto \Gamma(\alpha + 1)\alpha/x^{1+\alpha}$.

3.6.2 *Pareto Distribution, Zipf–Mandelbrot and Bradford Laws*

For large enough values of the total number of scientists and the total number of publications, we can make the transition from discrete to continuous representation of the corresponding variables and laws. The continuous analog of Lotka law, (3.66), is the Pareto distribution

$$p(x) = \frac{\alpha}{x_0}\left(\frac{x_0}{x}\right)^{\alpha+1}; \qquad x \geq x_0; \quad \alpha > 0 \tag{3.71}$$

which describes the distribution density for a number of scientists with x papers; x_0 is the minimal productivity $x_0 << x << \infty$, a continuous quantity.

Zipf law is connected to the principle of least effort (Zipf 1949): a person will try to solve his problems in such a way as to minimize the total work that he must do in the solution process. For example, to express with many words what can be expressed with a few is meaningless. Thus, it is important to summarize an article using a small number of meaningful words. Bradford law for the scattering of articles over different journals is connected to the success-breeds-success (SBS) principle (Price 1976): success in the past increases chances for some success in the future. For example, a journal that has been frequently consulted for some purpose is more likely to be read again, rather than one of previously infrequent use.

In order to obtain the law of Zipf–Mandelbrot, we start from the following version of Lotka law : $n_x = C/(1+x)^{1+\alpha}$, where x is the scientist's productivity, α is a characteristic exponent, C is a constant which in most cases is equal to the number of authors with the minimal productivity $x = 1$, i.e., to n_1. On the basis of this formula, the number of scientists r who are characterized by productivity $x_r < x < k_{max}$ (k_{max} is the maximal productivity of a scientist) reads

$$r = \sum_{x=x_r}^{k_{max}} n_r \approx C\int_{x_r}^{k_{max}} \frac{dx}{x^{1+\alpha}} = \frac{C}{\alpha}\left(\frac{1}{x_r^{\alpha}} - \frac{1}{k_{max}^{\alpha}}\right). \tag{3.72}$$

Depending on the value of x_r, r can have values $1, 2, 3, \dots$ and in such a way the scientists can be ranked. If all scientists of a scientific community working on the same topic are ranked in the order of the decrease of their productivity, the place of a scientist who has written x_r papers will be determined by his/her rank r. When the productivity of a scientist x_r is found from (3.72) as a function of rank r, the relationship

$$x_r = \left(\frac{A}{r+B}\right)^{\gamma}; \quad A = (C/\alpha)^{1/\alpha}; \quad B = C/(\alpha k_{\max}^{\alpha}); \quad \gamma = 1/\alpha. \tag{3.73}$$

This is the rank law of Zipf–Mandelbrot, which generalizes Zipf law: $f(r) = cr^{-\beta}$; $r = 1, 2, 3, \dots$, where c and β are parameters. Zipf law was discovered by counting words in books. If words in a book are ranked in decreasing order according to their number of occurrences, then Zipf law states that the number of occurrences of a word is inversely proportional to its rank r.

Assuming that in Lotka law the exponent takes the value $\alpha = 1$ and that in most cases $C = n_1$, one has $x_r = n_1/(r+a)$, where $a = n_1/k_{max}$, $r \geq 0$. Integration of the last relationship yields the total productivity $R(n)$ of all scientists, beginning with the one with the greatest productivity k_{max} and ending with the scientist whose productivity corresponds to the rank n (the scientists are ranked in the order of diminishing productivity; the rank is assumed to be a continuous-like variable):

$$R(n) = n_1 \ln\left(\frac{n}{a} + 1\right). \tag{3.74}$$

This is Bradford law. According to this law, for a given topic, a large number of relevant articles will be concentrated in a small number of journals. The remaining articles will be dispersed over a large number of journals. Thus, if scientific journals are arranged in order of decreasing published articles on a given subject, they may be split to a core of journals more particularly devoted to the subject and a shell consisting of sub-shells of journals containing the same numbers of articles as the core. Then the number of journals from the core zone and succeeding sub-shells will follow the relationship $1 : n : n^2 : \ldots.$

Key point Nr. 15
The Zipf–Pareto law, in the case of the distribution of scientists with respect to their productivity, indicates that one can always single out a small number of productive scientists who wrote the greatest number of papers on a given subject, and a large number of scientists with low productivity. The same applies also to scientific contacts, citation networks, etc. This specific feature (so-called hierarchical stratification) of the Zipf–Pareto law reflects a basic mechanism in the formation of stable complex systems. This can/must be taken into account in the process of planning and the organization of science.

3.7 Concluding Remarks

Knowledge has a complex nature. It can be created. It can lead to innovations and new technologies, and on this base, knowledge supports the advance and economic growth of societies. Knowledge can be collected. Knowledge can be spread. Diffusion of ideas is closely connected to the collection and spreading of knowledge. Some stages of the diffusion of ideas can be described by epidemic models of scientific and technological systems. Most of the models described here are deterministic, but if the internal and external fluctuations are strong, then different kinds of models can be applied taking into account stochastic features.

Much information about properties and stability of the knowledge systems can be obtained by the statistical approach on the basis of distributions connected to the Lotka–Volterra models of diffusion of knowledge. Interestingly, new terms occur in the usual evolution equations because of the variability and flexibility in the opinions of actors, due to media contacts or interpersonal contacts, when exchanging ideas.

The inclusion of spatial variables in the models leads to new research topics, such as questions on the spreading of systems of ideas and competition among ideas in different areas/countries.

In conclusion, the epidemiological perspective renders a piece of mosaic to a better understanding of the dynamics of diffusion of ideas in science, technology, and society, which should be one of the main future tasks of the science of science (Wagner-Döbler and Berg 1994).

Acknowledgements Thanks to the editors of the book for inspiring us into writing this work. The authors gratefully acknowledge stimulating discussions with many wonderful colleagues at several meetings of the ESF Action COST MP-0801 'Physics of Competition and Conflict'.

References

Allen JC (1975) Mathematical model of species interactions in time and space. Am Nat 109(967):319–342 (DOI:10.1086/283000, stable JSTOR URL: http://www.jstor.org/stable/2459697)

Amabile TM, Conti R, Coon H, Lazenby J, Herron M (1996) Assessing the work environment for creativity. Acad Manage Rev 39(5):1154–1184 (DOI:10.2307/256995)

Anderson RM, May RM (eds) (1982) Population biology of infectious diseases: Report of the Dahlem workshop on population biology of infectious disease agents, Berlin 1982, March 14–19. Dahlem Workshop Reports. Life Sciences Research Reports, vol 25. Springer, Berlin

Antonelli C (1996) Localized knowledge percolation processes and information networks. J Evol Econ 6(3):281–295 (DOI:10.1007/BF01193634)

Ausloos M, Lambiotte R, Scharnhorst A, Hellsten I (2008) Andrzej Pękalski networks of scientific interests with internal degrees of freedom through self-citation analysis. Int J Mod Phys C 19(3):371–384 (DOI:10.1142/S0129183108012224), also available as arXiv preprint arXiv:0710.1800

Ausloos M (2010) On religion and language evolutions seen through mathematical and agent based models. In: Rangacharyulu C, Haven E (eds) Proceedings of the first interdisciplinary CHESS interactions conference: Saskatoon, Saskatchewan, Canada, 17–20 August 2009. World Scientific, Singapore, pp 157–182 (DOI:10.1142/9789814295895_0009), also available as arXiv preprint arXiv:1103.5382

Barro RJ, Sala-i-Martin X (2004) Economic growth, 2nd edn. MIT Press, Cambridge, MA

Bartholomew DJ (1982) Stochastic models for social processes, 3rd edn. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, Chichester

Becker GS, Murphy KM (1988) A theory of rational addiction. J Polit Econ 96(4):675–700 (DOI:10.1086/261558, stable JSTOR URL: http://www.jstor.org/stable/1830469)

Becker GS (1996) Accounting for tastes. Harvard University Press, Cambridge, MA

Bernius S (2010) The impact of open access on the management of scientific knowledge. Online Inf Rev 34(4):583–603 (DOI:10.1108/14684521011072990)

Bettencourt LMA, Cintrón-Arias A, Kaiser DI, Castillo-Chávez C (2006) The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. Physica A 364:513–536 (DOI:10.1016/j.physa.2005.08.083), also available as arXiv preprint arXiv:physics/050206

Bettencourt LMA, Kaiser DI, Kaur J, Castillo-Chávez C, Wojick DE (2008) Population modeling of the emergence and development of scientific fields. Scientometrics 75(3):495–518 (DOI:10.1007/s11192-007-1888-4)

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Phys Rep 424(4–5):176–308 (DOI:10.1016/j.physrep.2005.10.00)

Bourdieu P (1986) Forms of capital. In: Richardson JG (ed) Handbook of theory and research for the sociology of education. Greenwood, New York, NY, pp 241–258

Braun T, Glänzel W, Schubert A (1985) Scientometric indicators: A 32-country comparative evaluation of publishing performance and citation impact. World Scientific, Singapore

Brauer F, Castillo-Chavez C (2001) Mathematical models in population biology and epidemiology. Texts in Applied Mathematics, vol 40. Springer, New York, NY

Bruckner E, Ebeling W, Scharnhorst A (1989) Stochastic dynamics of instabilities in evolutionary systems. Syst Dyn Rev 5(2):176–191 (DOI:10.1002/sdr.4260050206)

Bruckner E, Ebeling W, Scharnhorst A (1990) The application of evolution models in scientometrics. Scientometrics 18(1):21–41 (DOI:10.1007/BF02019160)

Bruckner E, Ebeling W, Jiménez-Montaño MA, Scharnhorst A (1996) Nonlinear stochastic effects of substitution – an evolutionary approach. J Evol Econ 6(1):1–30 (DOI:10.1007/BF01202370)

Bryman A (1988) Quantity and quality in social research. Contemporary Social Research Series, vol 18. Unwin Hyman, London

Burrell QL (2007) Hirsch's h-index: A stochastic model. J Informetr 1(1):16–25 (DOI:10.1016/j.joi.2006.07.001)

Casetti E, Semple RK (1969) Concerning the testing of spatial diffusion hypotheses. Geogr Anal 1(3):254–259 (DOI:10.1111/j.1538-4632.1969.tb00622.x)

Castiaux A (2007) Radical innovation in established organizations: Being a knowledge predator. J Eng Tech Manag 24(1–2):36–52 (DOI:10.1016/j.jengtecman.2007.01.003)

Chen C, Hicks D (2004) Tracing knowledge diffusion. Scientometrics 59(2):199–211 (DOI:10.1023/B:SCIE.0000018528.59913.48)

Chen YS, Cheng PP, Tong Y (1993) Theoretical foundation of the 80/20 rule. Scientometrics 28(2):183–204 (DOI:10.1007/BF02016899)

Chung KH, Cox RAK (1990) Patterns of productivity in the finance literature: A study of the bibliometric distributions. J Finance 45(1):301–309 (DOI:10.2307/2328824, stable JSTOR URL: http://www.jstor.org/stable/2328824)

Coleman JC (1988) Social capital in the creation of human capital. Am J Sociol 94(Suppl):S95–S120. (DOI:10.1086/228943, stable JSTOR URL: http://www.jstor.org/stable/2780243)

Cowan R, Foray D (1997) The economics of codification and the diffusion of knowledge. Ind Corp Change 6(3):595–622 (DOI:10.1093/icc/6.3.595)

Dahlman C, Zhihua Zeng D, Wang S (2007) Enhancing China's competitiveness through life long learning. The World Bank, Washington, DC

Dahlman C (2009) Different innovation strategies, different results: Brazil, Russia, India, China and Korea (the BRICKs). In: Chandra V, Eröcal D, Padoan PC, Primo Braga AC (eds) Innovation and growth: Chasing a moving frontier. OECD, Paris; International Bank for Reconstruction and Development/The World Bank, Washington, DC, pp 131–168, available online at the URL: http://www.oecd.org/document/35/0,3343,en_2649_37417_44268835_1_1_1_37417,00.html

Daley DJ (1967) Concerning the spread of news in a population of individuals who never forget. Bull Math Biol 29(2):373–376 (DOI:10.1007/BF02476908)

Davis JB (2003) The theory of the individual in economics: Identity and value. Routledge Advances in Social Economics. Routledge, London (DOI:10.4324/9780203457689)

Dietz K (1967) Epidemics and rumours: A survey. J R Stat Soc Ser A 130(4):505–528 (DOI:10.2307/2982521, stable JSTOR URL: http://www.jstor.org/stable/2982521)

Dimitrova ZI, Vitanov NK (2000) Influence of adaptation on the nonlinear dynamics of a system of competing populations. Physc Lett A 272(5–6):368–380 (DOI:10.1016/S0375-9601(00)00455-2)

Dimitrova ZI, Vitanov NK (2001) Adaptation and its impact on the dynamics of a system of three competing populations. Physica A 300(1–2):91–115 (DOI:10.1016/S0378-4371(01)00330-2)

Dimitrova ZI, Vitanov NK (2001) Dynamical consequences of adaptation of the growth rates in a system of three competing populations. J Phys A Math Gen 34(37):7459–7473 (DOI:10.1088/0305-4470/34/37/303)

Dimitrova ZI, Vitanov NK (2004) Chaotic pairwise competition. Theor Popul Biol 66(1):1–12 (DOI:10.1016/j.tpb.2003.10.008)

Dodd SC (1958) Formulas for spreading opinions. Public Opin Q 22(4):537–554 (DOI:10.1086/266827, stable JSTOR URL: http://www.jstor.org/stable/2746601)

Dolfsma W (2008) Knowledge economies: Organization, location and innovation. Routledge Studies in Global Competition, vol 39. Routledge, London

Ebeling W, Engel A, Mazenko VG (1986) Modeling of selection processes with age-dependent birth and death rates. BioSystems 19(3):213–221 (DOI:10.1016/0303-2647(86)90040-7)

Ebeling W, Engel A, Feistel R (1990) Physik der Evolutionsprozesse. Akademie-Verlag, Berlin

Ebeling W, Feistel R, Hartmann-Sonntag I, Schimansky-Geier L, Scharnhorst A (2006). New species in evolving networks – stochastic theory of sensitive networks and applications on the metaphorical level. BioSystems 85(1):65–71 (DOI:10.1016/j.biosystems.2006.02.013)

Ebeling W, Scharnhorst A (1985) Selforganization models for field mobility of physicists. Czech J Phys 36(1):43–46 (DOI:10.1007/BF01599723)

Ebeling W, Scharnhorst A (2000) Evolutionary models of innovation dynamics. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) Traffic and granular flow' 99. Social, traffic and granular dynamics. Springer, Berlin, pp 43–56

Edelstein-Keshet L (1988) Mathematical models in biology. The Random House/Birkhäuser Mathematics Series. Random House, New York, NY

Egghe L (1998) Mathematical theories of citation. Scientometrics 43(1):57–62 (DOI:10.1007/BF02458394)

Egghe L, Rousseau R (1990) Introduction to informetrics: Quantitative methods in library, documentation and information science. Elsevier, Amsterdam

Epstein JM (1997) Nonlinear dynamics, mathematical biology, and social science. Santa Fe Institute Studies in the Sciences of Complexity. Lecture Note, vol 4. Addison-Wesley, Reading, MA

Fernández-Camo A, Toralbo M, Vallejo M (2004) Reconsidering Price's model of scientific growth: An overview. Scientometrics 61(3):301–321 (DOI:10.1023/B:SCIE.0000045112.11562.11)

Foray D (2004) The economics of knowledge. MIT Press, Cambridge, MA. Revised and extended translation of: Foray D (2000) L'conomie de la connaissance. La Decouverte, Paris

Gao X, Guan J (2009) A scale-independent analysis of the performance of the Chinese innovation system. J Informetr 3(4):321–331 (DOI:10.1016/j.joi.2009.04.004)

Gardiner CW (1983) Handbook of stochastic methods for physics, chemistry and the natural sciences. Springer Series in Synergetics, vol 13. Springer, Berlin

Gause GF (1934) The struggle for existence. Williams and Wilkins, Baltimore, MD

Geroski PA (2000) Models of technology diffusion. Res Policy 29(4):603–625 (DOI:10.1016/S0048-7333(99)00092-X)

Gilbert GN (1978) Measuring the growth of science: A review of indicators of scientific growth. Scientometrics 1(1):9–34 (DOI:10.1007/BF02016837)

Glänzel W, Schoepflin U (1994) A stochastic model for the ageing of scientific literature. Scientometrics 30(1):49–64 (DOI:10.1007/BF02017212)

Goffman W (1966) Mathematical approach to the spread of scientific ideas – the history of mast cell research. Nature 212(5061):449–452 (DOI:10.1038/212449a0)

Goffman W, Newill VA (1964) Generalization of epidemic theory: An application to the transmission of ideas. Nature 204(4955):225–228 (DOI:10.1038/204225a0)

Haitun SD (1982) Stationary scientometric distributions. Part II. Non-Gaussian nature of scientific activities. Scientometrics 4(2):89–104 (DOI:10.1007/BF02018448)

Hänggi P, Thomas H (1982) Stochastic processes: Time evolution, symmetries and linear response. Phys Rep 88(4):207–319 (DOI:10.1016/0370-1573(82)90045-X)

Hellsten I, Lambiotte R, Scharnhorst A, Ausloos M (2006) A journey through the landscape of physics and beyond – the self-citation patterns of Werner Ebeling. In: Pöschel T, Malchow H, Schimansky–Geier L (eds) Irreversible Prozesse und Selbstorganisation. Logos Verlag, Berlin, pp 375–384

Hellsten I, Lambiotte R, Scharnhorst A, Ausloos M (2007) Self-citations, co-authorships and keywords: A new method for detecting scientists field mobility? Scientometrics 72(3):469–486 (DOI:10.1007/s11192-007-1680-5)

Hellsten I, Lambiotte R, Scharnhorst A, Ausloos M (2007) Self-citation networks as traces of scientific careers. In: Torres-Salinas D, Moed H (eds) Proceedings of the ISSI 2007, 11th International Conference of the International Society for Scientometrics and Informetrics, CSIC, Madrid, Spain, June 25–27, 2007. CINDOC-CSIC, Madrid, vol 1, pp 361–367

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569–16572 (DOI:10.1073/pnas.0507655102, stable JSTOR URL: http://www.jstor.org/stable/4152261), also available as arXiv preprint arXiv:physics/0508025

Howells JRL (2002) Tacit knowledge, innovation and economic geography. Urban Stud 39(5–6):871–884 (DOI:10.1080/00420980220128354)

Ivanova K, Ausloos M (1999) Application of the detrended fluctuation analysis (DFA) method for describing cloud breaking. Physica A 274(1–2):349–354 (DOI:10.1016/S0378-4371(99)00312-X)

Jaffe AB (1986) Technological opportunity and spillovers of R&D: Evidence from firms patents, profits, and market value. Am Econ Rev 76(5):984–1001 (stable JSTOR URL: http://www.jstor.org/stable/1816464), also availabe online as National Bureau of Economic Research (NBER) Working Paper No. 1815 at the URL: http://www.nber.org/papers/w1815

Jaffe AB, Trajtenberg M (eds) (2002) Patents, citations and innovations: A window on the knowledge economy. MIT Press, Cambridge, MA

Karmeshu (1982) Time lag in a diffusion model of information. Math model 3(2):137–141 (DOI:10.1016/0270-0255(82)90018-5)

Katz JS (1999) The self-similar science system. Res Policy 28(5):501–517 (DOI:10.1016/S0048-7333(99)00010-4)

Kealey T (2000) More is less. Economists and governments lag decades behind Derek Price's thinking. Nature 405(6784):279–279 (DOI:10.1038/35012717)

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. Proc R Soc Lond A Math Phys Character 115(772):700–721 (DOI:10.1098/rspa.1927.0118, stable JSTOR URL: http://www.jstor.org/stable/94815)

Kerner EH (1959) Further considerations on the statistical mechanics of biological associations. Bull Math Biol 21(2):217–253 (DOI:10.1007/BF02476361)

Keynes JM (1930) A treatise on money. 2 Volumes. Harcourt, Brace and Co., New York, NY. Reprinted in: Keynes JM (1971, 1989, 2nd edn) The collected writings of John Maynard Keynes: Volume V: A treatise on money: In two volumes: 1. The pure theory of money, Volume VI: A treatise on money: In two volumes: 2. The applied theory of money. Macmillan, London

Kiss IZ, Broom M, Graze PG, Rafols I (2000) Can epidemic models describe the diffusion of topics across disciplines. J Infometr 4(1):74–82 (DOI:10.1016/j.joi.2009.08.002), also available as arXiv preprint arXiv:0905.3585

Kiss IZ, Green DM, Kao RR (2005) Disease contact tracing in random and clustered networks. Proc R Soc B Biol Sci 272(1570):1407–1414 (DOI:10.1098/rspb.2005.3092, stable JSTOR URL: http://www.jstor.org/stable/30047699)

Kot SM (1987) The stochastic model of evolution of scientific disciplines. Scientometrics 12(3–4):197–205 (DOI:10.1007/BF02016292)

Kucharavy D, Schenk E, de Guio R (2009) Long-run forecasting of emerging technologies with logistic models and growth of knowledge. In: Roy R, Shehab E (eds) Proceedings of the 19th CIRP design conference – competitive design, Cranfield, UK. Cranfield University, Cranfield, pp 277–284, available online at the URL: http://dspace.lib.cranfield.ac.uk/handle/1826/3730

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago, IL

Lal VB, Karmeshu, Kaicker S (1988) Modeling innovation diffusion with distributed time lag. Technol Forecast Soc Change 34(2):103–113 (DOI:10.1016/0040-1625(88)90060-1)

Leydesdorff L (2006) The knowledge-based economy: Modeled, measured, simulated. Universal Publishers, Boca Raton, FL

Li W (2002) Zipf's law everywhere. Glottometrics 5:15–41, available online at the URL: http://www.nslij-genetics.org/wli/pub/glottometrics02.pdf

Lotka AJ (1912) Ein Fall von Autokatakinese mit oscillatorischem Verlauf. Zeitschrift für Physikalische Chemie, Stöchiometrie und Verwandtschaftslehre 80(2):159–164

Lotka AJ (1925) Elements of physical biology. Williams & Wilkins Company, Baltimore, MD, available online at the URL: http://www.archive.org/details/elementsofphysic017171mbp. Reprinted as: Lotka AJ (1956) Elements of mathematical biology. Dover Books on the Biological Sciences. Dover Publications, New York, NY

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–323

Ma Z, Li J (eds) (2009) Dynamical modeling and analysis of epidemics. World Scientific, Singapore, available online at the URL: http://ebooks.worldscinet.com/ISBN/9789812797506/9789812797506.html

Magyari-Beck I (1984) A method of measurement of scientific production quality. Science of Science 4(2):183–195

Mahajan V, Peterson RA (1985) Models for innovation diffusion. Quantitative Applications in the Social Sciences, vol 48. Sage Publications, Beverly Hills, CA

Mansfield E (1961) Technical change and the rate of imitation. Econometrica 29(4):741–766 (DOI:10.2307/1911817, stable JSTOR URL: http://www.jstor.org/stable/191181)

Marshall A (1920) Principles of economics: An introductory volume, 8th edn. McMillan, London, available online ((1907) 5th edn) at the URL: http://www.archive.org/details/principlesofecon01marsuoft

May RM (1974) Stability and complexity in model ecosystems, 2nd edn. Monographs in Population Biology, vol 6. Princeton University Press, Princeton, NJ

Meade N, Islam T (1995) Forecasting with growth curves: An empirical comparison. Int J Forecast 11(2):199–215 (DOI:10.1016/0169-2070(94)00556-R)

Meyer PS (1994) Bi-logistic growth. Technol Forecast Soc Change 47(1):89–102 (DOI:10.1016/0040-1625(94)90042-6)

Meyer PS, Yung JW, Ausubel JH (1999) A primer on logistic growth and substitution: The mathematics of the Loglet Lab software. Technol Forecast Soc Change 61(3):247–271 (DOI:10.1016/S0040-1625(99)00021-9)

Modis T (2003) A scientific approach to managing competition. Ind Physicist 9(1):24–27, available online at the URL: http://www.aip.org/tip/INPHFA/vol-9/iss-1/p25.htm

Modis T (2007) Strengths and weaknesses of S-curves. Technol Forecast Soc Change 74(6):866–872 (DOI:10.1016/j.techfore.2007.04.005)

Morone P, Taylor R (2010) Knowledge diffusion and innovation: Modelling complex entrepreneurial behaviours. Edward Elgar Publishing Inc., Northampton, MA

Murray JD (1989) Mathematical biology. Biomathematics, vol 19. Springer, Berlin

Nonaka I (1994) A dynamic theory of organizational knowledge creation. Organ Sci 5(1):14–37 (DOI:10.1287/orsc.5.1.14, stable JSTOR URL: http://www.jstor.org/stable/2635068)

Nonaka I, Takeuchi H (1995) The knowledge creating company: How Japanese companies create the dynamics of innovation. Oxford University Press, Oxford

Nonaka I, Konno N (1998) The concept of "Ba": Building a foundation for knowledge creation. Calif Manage Rev 40(3):40–54 (DOI:10.1225/CMR107)

Nowakowska M (1973) Epidemical spread of scientific objects: An attempt of empirical approach to some problems of meta-science. Theory Decis 3(3):262–297 (DOI:10.1007/BF00139506)

Noyons ECM, Van Raan AFJ (1998) Advanced mapping of science and technology. Scientometrics 41(1–2):61–67 (DOI:10.1007/BF02457967)

Odum EP (1959) Fundamentals of ecology. W.B. Saunders, Philadelphia, PA

Okubo A (1980) Diffusion and ecological problems: Mathematical models. Biomathematics, vol 10. Springer, Berlin. Extended translation of: Okubo A (1975) *Seitaigaku to kakusan*. Tukijishokan, Tokyo

Plesk PE (1997) Creativity, innovation and quality. ASQ Quality Press, Milwaukee, WI

Price DJ de Solla (1951) Quantitative measures of the development of science. Archives Internationale d'Histoire de Sciences 4(14):86–93, available online at the URL: http://garfield.library.upenn.edu/price/pricequantitativemeasures1951.pdf

Price DJ de Solla (1956) The exponential curve of science. Discovery 17(1):240–243. Reprinted in: Barber B, Hirsch W (eds) (1962) The sociology of science. The Free Press of Glencoe, New York, NY, pp 516–524

Price DJ de Solla (1961) Science since Babylon. Yale University Press, New Haven, CT

Price DJ de Solla (1963) Little science, big science. Columbia University Press, New York, NY

Price DJ de Solla (1971) Principles for projecting funding of academic science in the 1970s. Sci Stud 1(1):85–94 (DOI:10.1177/030631277100100106, stable JSTOR URL: http://www.jstor.org/stable/370198)

Price DJ de Solla (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306 (DOI:10.1002/asi.4630270505), also available online at the URL: http://www.asis.org/Publications/JASIS/Best_Jasist/1976pricejasistarticle.pdf

Price DJ de Solla, Gürsey S (1975) Some statistical results for the numbers of authors in the states of the United States and the nations of the world. In: Who is publishing in science 1975 annual. Institute for Scientific Information, Philadelphia, PA, pp 26–34

Putnam RD, Leonardi R, Nanetti RY (1993) Making democracy work: Civic transitions in modern Italy. Princeton University Press, Princeton, NJ

Risken H (1984) The Fokker-Planck equation: Methods of solution and applications. Springer Series in Synergetics, vol 18. Springer-Verlag, Berlin

Rogers E (1962) Diffusion of innovations. The Free Press, New York, NY; Collier-MacMillan, London

Romanov AK, Terekhov AI (1997) The mathematical model of productivity- and age-structured scientific community evolution. Scientometrics 39(1):3–17 (DOI:10.1007/BF02457427)

Romer PM (2002) When should we use intellectual property rights? Am Econ Rev 92(2):213–216 (DOI:10.1257/000282802320189276, stable JSTOR URL: http://www.jstor.org/stable/3083404). The article is part of: Baldwin JE, Oaxaca RL (eds) Papers and proceedings of the one hundred fourteenth annual meeting of the American Economic Association, Atlanta, GA, January 4–6, 2002. Am Econ Rev 92(2):1–478

Romer PM (1994a) The origins of endogeneous growth. J Econ Perspect 8(1):3–22 (stable JSTOR URL: http://www.jstor.org/stable/2138148)

Romer PM (1994b) New goods, old theory, and the welfare costs of trade restrictions. J Dev Econ 43(1):5–38 (DOI:10.1016/0304-3878(94)90021-3), also availabe online as National Bureau of Economic Research (NBER) Working Paper No. 4452 at the URL: http://www.nber.org/papers/w4452

Romer D (1996) Advanced macroeconomics. McGraw-Hill, New York, NY

Saviotti PP (1999) Knowledge, information and organisational structures. In: Robertson PL (ed) Authority and control in modern industry: Theoretical and empirical perspectives. Routledge Studies in Business Organization and Networks, vol 10. Routledge, London, pp 120–139 (DOI:10.4324/9780203435403.ch4)

Scharnhorst A (1998) Citation networks, science landscapes and evolutionary strategies. Scientometrics 43(1):95–106 (DOI:10.1007/BF02458399)

Scharnhorst A (2001) Constructing knowledge landscapes within the framework of geometrically oriented evolutionary theories. In: Matthies M, Malchow H, Kriz J (eds) Integrative systems approaches to natural and social dynamics. Springer, Berlin, pp 505–515

Sharif MN, Ramanathan K (1981) Binomial innovation diffusion models with dynamic potential adopter population. Technol Forecast Soc Change 20(1):63–87 (DOI:10.1016/0040-1625(81)90041-X)

Small H (1997) Update on science mapping: Creating large document spaces. Scientometrics 38(2):275–293 (DOI:10.1007/BF02457414)

Small H (1998) A general framework for creating large-scale maps of science in two or three dimensions: The SciViz system. Scientometrics 41(1–2):125–133 (DOI:10.1007/BF02457973)

Small H (2006) Tracking and predicting growth areas in science. Scientometrics 68(3):595–610 (DOI:10.1007/s11192-006-0132-y)

AQ3

Soler JM (2007) A rational indicator of scientific creativity. J Informetr 1(2):123–130 (DOI:10.1016/j.joi.2006.10.004), also available as arXiv preprint arXiv:physics/0608006

Solomon S, Richmond P (2001) Power laws of wealth, market order volumes and market returns. Physica A 299(1–2):188–197 (DOI:10.1016/S0378-4371(01)00295-3), also available as arXiv preprint arXiv:cond-mat/0102423

Solomon S, Richmond P (2002) Stable power laws in variable economics; Lotka–Volterra implies Pareto–Zipf. Eur Phys J B 27(2):257–261 (DOI:10.1140/epjb/e20020152)

Stiglitz JE (1987) Learning to learn, localized learning and technological progress. In: Dasgupta P, Stoneman P (eds) Economic policy and technological performance. Cambridge University Press, Cambridge, pp 125–153

Szydlowski M, Krawiez A (2001) Scientific cycle model with delay. Scientometrics 52(1):83–95 (DOI:10.1023/A:1012751028630)

Szydlowski M, Krawiez A (2009) Growth cycles of knowledge. Scientometrics 78(1):99–111 (DOI:10.1007/s11192-007-1958-7)

Vandewalle N, Ausloos M (1997) Coherent and random sequences in financial fluctuations. Physica A 246(3–4):454–459 (DOI:10.1016/S0378-4371(97)00366-X)

van Kampen NG (1981) Stochastic processes in physics and chemistry. North-Holland Publishing Company, Amsterdam

van Raan AFJ (1997) Scientometrics: State of art. Scientometrics 38(1):205–218 (DOI:10.1007/BF02461131)

van Raan AFJ (2000) On growth, ageing and fractal differentiation of science. Scientometrics 47(2):347–362 (DOI:10.1023/A:1005647328460)

Vitanov NK, Dimitrova ZI, Kantz H (2006) On the trap of extinction and its elimination. Phys Lett A 349(5):350–355 (DOI:10.1016/j.physleta.2005.09.043)

Vitanov NK, Dimitrova ZI, Ausloos M (2010) Verhulst-Lotka–Volterra (VLV) model of ideological struggle. Physica A 389(21):4970–4980 (DOI:10.1016/j.physa.2010.06.032), also available as arXiv preprint arXiv:1103.5362

Vitanov NK, Jordanov IP, Dimitrova ZI (2009) On nonlinear population waves. Appl Math Comput 215(8):2950–2964 (DOI:10.1016/j.amc.2009.09.041)

Vitanov NK, Jordanov IP, Dimitrova ZI (2009) On nonlinear dynamics of interacting populations: Coupled kink waves in a system of two populations. Commun Nonlinear Sci Numer Simul 14(5):2379–2388 (DOI:10.1016/j.cnsns.2008.07.015)

Volterra V (1927) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Memorie della Classe di Scienze Fisiche, Matematiche e Naturali, Accademia Nazionale dei Lincei, Roma, 6th series, 2(3):31–113. English translation: Variations and fluctuations of the number of individuals in animal species living together. In: Chapman RN (1931) Animal ecology. McGraw Hill, New York, NY, pp 409–448

Wagner-Döbler R, Berg J (1994) Regularity and irregularity in the development of scientific disciplines: The case of mathematical logics. Scientometrics 30(1):303–319 (DOI:10.1007/BF02017230)

Walsh JR (1935) Capital concept applied to man. Q J Econ 49(2):255–285 (DOI:10.2307/1884067, stable JSTOR URL: http://www.jstor.org/stable/1884067)

Willson WG, de Roos AM (1993) Spatial instabilities within the diffusive Lotka–Volterra system: Individual-based simulation results. Theor Popul Biol 43(1):91–127 (DOI:10.1006/tpbi.1993.1005)

Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Jones DF (ed) Proceedings of the Sixth International Congress of Genetics. Brooklyn Botanic Garden, Brooklyn, NY, part 1, pp 356–366, available online at the URL: http://www.esp.org/books/6th-congress/facsimile/contents/6th-cong-p356-wright.pdf

Yablonsky AI (1980) On fundamental regularities of the distribution of scientific productivity. Scientometrics 2(1):3–34 (DOI:10.1007/BF02016597)

Yablonsky AI (1985) Stable non-Gaussian distributions in scientometrics. Scientometrics 7(3–6):459–470 (DOI:10.1007/BF02017161)

Ziman JM (1969) Information, communication, knowledge. Nature 224(5217):318–324 (DOI:10.1038/224318a0)

Zipf GK (1949) Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley, Cambridge, MA

AUTHOR QUERIES

AQ1. First author has been considered as corresponding author. Please check.
AQ2. Please check whether the inserted citation for Tables 3.1–3.5 are appropriate.
AQ3. Please cite ref Small (2006) in text.

Chapter 4
Agent-Based Models of Science

Nicolas Payette

4.1 What are Agent-Based Models?

This first section is mostly an introduction to ABMs in general. We will first take a look at where they come from and what their main characteristics are. We will then bring forward a few methodological considerations and illustrate some of those with an actual agent-based model of science (Table 4.1). AQ1

4.1.1 A Little History

Agent-based models are intimately linked with computers and, perhaps unsurprisingly, we count John Von Neumann as a pioneer of both. In the late 1940s, Von Neumann (with an eye towards artificial intelligence) was interested in self-reproducing, self-regulating systems. Inspired by ideas from colleague Stanisław Ulam, he designed the first cellular automaton.

What he came up with is a system made of "cells" laid out on a discrete, orthogonal, grid (later described in von Neumann 1966). Time, in the system, is also discrete, and at each time step, every cell updates its state according to a set of rules based on its previous state and the state of its neighbors on the grid. Each cell is a simple finite state machine, but the overall behavior of the system can become quite complex. Von Neumann used that framework to design what he called a "Universal Constructor": a pattern of cells that can reproduce itself over time, thereby providing a striking example of how an important system-level property (self-reproduction) can be achieved through the interaction of individual parts that behave independently from the whole (Fig. 4.1). AQ2

N. Payette (✉)
Université du Québec à Montréal, C.P. 8888, succ. Centre-ville, Montréal, QC, Canada H3C 3P8
e-mail: nicolaspayette@gmail.com

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_4,

Table 4.1 Major questions and their answers

Model	Major question(s) the model aims to answer	Key answers/insights in lay terms
Gilbert (1997)	Is it possible to generate some of the quantitative features of science by using simple mechanisms, and if so, what are those?	"it is possible to generate many of the quantitative features of the present structure of science and that one way of looking at scientific activity is as a system in which scientific papers generate further papers, with authors (scientists) playing a necessary but incidental role."
Edmonds (2007)	What can we learn by modeling the collective scientific process as a form of distributed computing?	The collective scientific process, modeled as a distributed theorem prover, "has the potential to [serve as an] intermediate between observations concerning how science works and areas of distributed knowledge discovery in computer science"
Zollman (2007)	What is the relation between the network structures of a community of scientists and its ability to converge on the right hypothesis given limited information?	A more connected network will converge much more rapidly on an hypothesis, but is much more likely to converge on the wrong hypothesis: there is an important trade-off between speed and accuracy.
Sun and Naveh (2009)	What is the relationship between individual cognitive factors and some of the quantitative features of the scientific system?	"while different cognitive settings may affect the aggregate number of scientific articles produced by the model, they do not generally lead to different distributions of number of articles per author. ... using more cognitively realistic models in simulations may lead to novel insights."
Weisberg and Muldoon (2009)	Which project selection strategies by individual researchers lead to the optimal distribution of cognitive labor for the scientific community?	"scientists need to really divide their cognitive labor, coordinating in such a way to take account of what other scientists are doing" and "a mixed strategy where some scientists are very conservative and others quite risk taking, leads to the maximum amount of epistemic progress in the scientific community."
Grim (2009)	What is the relation between the network structures of a community of scientists and its epistemic success in different epistemic landscapes?	Mean path length in the giant cluster of an epistemic network qualitatively matches the epistemic success of a community.
Muldoon and Weisberg (2010)	What is the effect of idealizations about the rationality of scientists on analytic models of the distribution of cognitive labor?	Analytic models of the distribution of cognitive labor are not robust against weakenings of idealizations about the rationality of scientists and the information available to them. Under certain conditions, this can lead to the model predicting outcomes that are qualitative opposites of the original model outcomes.

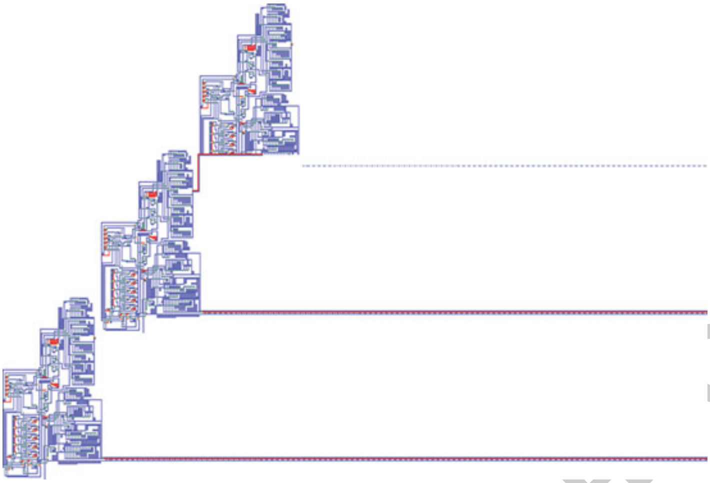


Fig. 4.1 Von Neumann's "Universal Constructor". Source: http://en.wikipedia.org/wiki/File:Nobili_Pesavento_2reps.png

What really brought cellular automata to the forefront, though, is mathematician John Conway's Game of Life (Gardner 1970). While von Neumann's cells could be in 29 different states and dozens of different rules were needed to describe transitions between them, Conway's cells (he called them "counters") are either "alive" or "dead." Only three rules are needed to describe their behavior:

1. Survivals. Every counter with two or three neighboring counters survives for the next generation.
2. Deaths. Each counter with four or more neighbors dies (is removed) from overpopulation. Every counter with one neighbor or none dies from isolation.
3. Births. Each empty cell adjacent to exactly three neighbors – no more, no fewer – is a birth cell. A counter is placed on it at the next move.

(Gardner 1970, p. 120)

These simple rules, when applied to different initial patterns of cells, give rise to an impressive (and well documented) menagerie of objects with complex behaviors: blinkers, toads, beacons, pulsars, gliders, guns, puffers, etc. Again, this shows how simple building blocks can be arranged in ways that lead to surprising (i.e., hard to predict) results (Fig. 4.2).

The systems we have seen so far are only models of very general phenomena ("life," self-replication), but the idea of cellular automata is also readily applicable to a lot of social phenomena. Notwithstanding debates around methodological individualism, many problems in the social sciences can be modeled as sets of individual agents locally interacting with each other in some explicit space.

The firsts of such models are Thomas Schelling's "Models of Segregation" (1969; 1971a; 1971b). In these, Schelling explores the mechanisms leading to the formation of clusters of homogeneous agents (i.e., ghettos) in geographical space.

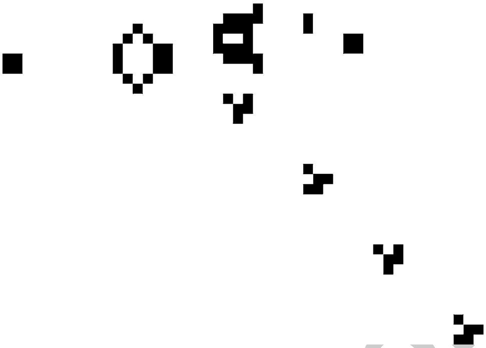


Fig. 4.2 Example of a complex pattern in Conway's Life. This is Bill Gosper's "glider gun": it builds "glider" patterns that move away from it. The gun keeps on generating gliders forever, unless it is disturbed by interference from another pattern. Source: http://en.wikipedia.org/wiki/File:Gospers_glider_gun.gif

Fig. 4.3 State of the Schelling neighborhood before the simulation: a mixed neighborhood

```
O # # # # O O O O     # #   O O
O   # O O O   # #     #   # O
#   # O O # #   #     O #   # #
#     # #     O # # # O O
O   O O # # # #     #   #   O O
# O #   O O #     O O     # #
# O O #         O O O # # #
O   # O   # #   # O O O       #
O   # O         # # O         #
O O       #     O # O O O O # #
O # # O O O O   O # #   O # #
#   O # O #   O O # O # O   O
O O     O # O # O O O     # #
```

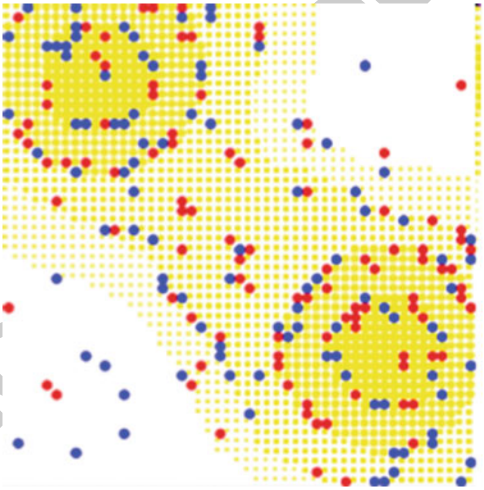
Space is modeled as a discrete grid, just like in von Neumann and Conway's automata, but this time, each individual cell represents a human agent. These agents can be either "stars" or "zeros" (taken to stand for different ethnicities), and they have preferences regarding the group membership of their neighbors on the grid. If they are not satisfied, they move to the closest location that satisfies their requirements. Schelling explored the dynamics of the model for many different initial patterns and many different distributions of preferences, but the general conclusion is that even with agents that have very high tolerance for neighbors different from themselves – just not wanting to be in too small a minority – segregation occurs consistently. As he points out himself, the *particular* outcome depends on details of a simulation run, but not the *character* of the outcome (Figs. 4.3 and 4.4).

Skipping far ahead, another milestone model is the much more complex Sugarscape (Epstein and Axtell 1996). While most other models were designed to investigate specific phenomena, the Sugarscape is a general framework for exploring a wide range of issues: biological and cultural evolution, trading, warfare, disease transmission, migration, pollution, etc. Agents in the Sugarscape are also situated on a grid, but this time the environment is not an empty container: it contains "sugar" and "spice," generic resources that the agents need to survive. Agents move around the grid, collecting these resources (which, afterwards, need to "grow back").

Fig. 4.4 State of the Schelling neighborhood after the simulation: segregation has occurred

```
# # # # #         # # # #   O O
# # # # # # # # # # # #   # O O
# # # # # # # # # # #   # # O O
#     # # # # # # # #     # O O
O O O O # # # # # # #   #
  O O O O O # # # O O   #   # #
  O O O O O O O O O O O # #
O     O   O O O   O O O       #
O     O             O         #
O O             O   O O O   # #
  O     O O O O   O     O O # #
    O   O     O O   O   O O
  O O     O   O   O O O     # #
```

Fig. 4.5 A view of the *Sugarscape*. Taken from Tony Bigbee's open-source reproduction of the Axtell and Epstein model using MASON (Bigbee et al. 2007)



this figure will be printed in b/w

Agents also differ from one another in more than group membership: they have different metabolic rates, vision, and life expectancy. These differences introduce interesting opportunities for interaction between agents. Take metabolic rate, for example: if you need more sugar and I need more spice, I can trade you sugar for spice.

While it is interesting in its own right to analyse the behavior of individual agents on the grid, it is the population-level patterns that are of most value to social science. For example, the individual wealth of the agents in the Sugarscape – the amount of resources they have accumulated – follows Pareto's Principle: a power-law distribution where very few agents control most of the wealth in the system. While Pareto's Principle has been observed in countless "real" social systems (starting with land ownership in early twentieth century Italy), Sugarscape is acknowledged to be the first computational generation of that pattern: it provides a set of micro-level mechanisms that are sufficient to generate that macro-level phenomenum. As we will see, the use of ABMs often follows that methodology (Fig. 4.5).

We chose to follow the historical path of cellular automata to introduce agent-based models, but other influences should also be acknowledged. Game theory (see, e.g., Axelrod and Hamilton 1981), artificial life (Reynolds 1987), connectionism (McClelland and Rumelhart 1987), genetic algorithms (Holland 1975) and artificial intelligence research in general also played important roles.

We opened the present section by stating that agent-based models are intimately linked with computers. While true, that statement can be slightly misleading: von Neumann's design for his Universal Constructor was not fully implemented until much later (Pesavento 1995), Conway designed Life on a Go board, and Schelling "ran" most of his simulations using pennies and dimes. In many cases, the local rules of behavior are simple enough that their results can be computed by hand. The computer is needed when the number of agents and steps in the simulation becomes overwhelming for the very limited computational resources of a human being.

The rise of computational resources in recent years has driven researchers to implement increasingly detailed models that aim to capture the finer aspects of social phenomena. A quick glance at the *Journal of Artificial Societies and Social Simulations* or at the "Model Archive" section of the *OpenABM* website[1] will reveal many of those, and there is also a trend to review and compare different classes of ABM (Cristelli et al. 2011).

4.1.2 Their Main Characteristics (and How They Apply to Models of Science)

Before paying attention to particular agent-based models of science, we want to say a few words about some general characteristics of ABMs. We will focus on the features listed by Joshua M. Epstein (2006): heterogeneity, autonomy, explicit space, local interactions and bounded rationality. These should not be taken as necessary conditions for a model to be considered agent-based. They should only be seen as establishing some kind of wittgensteinian family resemblance. They are not orthogonal either: some of them, such as local interactions and explicit space, for instance, overlap.

In this section, we will try to show that those features are well suited to the modeling of the scientific process.

Heterogeneity states that agents are not, as Epstein says, "aggregated in a few homogeneous pools" (2006, p. 6). Instead, they can differ from one another in as many ways as the parameter range for each of their individual properties will allow. While this is something that would be very hard to track with traditional analytical models, the computer makes it possible to deal with millions of heterogeneous agents.

---

[1] http://www.openabm.org/models/browse.

We can think of these varying properties as being either static or dynamic and either internal or external. Static properties are those that won't change through the agent's lifetime. It does not mean that they should be considered "innate," just that their value stays constant in the course of a simulation. Examples of such properties for scientists could be things like *creativity*, *communication skill*, *testing ability*, etc. Perhaps more interesting, though, are the dynamic properties of the agents: those that change, and hence, can be tracked through a simulation run. A dynamic property can be as simple as the amount of grant money a researcher currently has, but it can also be more than a simple numeric value: a list of the propositions that a scientist holds to be true, a memory of past interactions with other scientists, a current research goal, etc.

The examples that we have given so far are all internal properties. What we call external properties are relations between an agent and its environment. What university/lab/research center is a scientist attached to? Who are his collaborators? If space is represented, where is he? External properties are often dynamic but can also be static, depending on what the model is trying to capture.

Autonomy refers to the absence of central control. In the context of social simulation, this can be likened to a form of methodological individualism: while institutions (and other macro-structures) can set policies (rules, values, etc.) that will influence an agent's behaviour, they are not directly coordinating the agents or moving them around. At each time step in a simulation, agents make their own decisions in order to achieve their individual goals.

Explicit space requires that agents be situated in some environment. The behaviours available to an agent are partly determined by its position. In many ABMs, like in those we have seen so far, this is a grid representing geographic space, but it does not have to be. It can be something more abstract like (as we will see later) a scientist's position in an epistemic landscape or his position in a social network of collaboration. To quote Epstein again, "The main *desideratum* is that the notion of 'local' be well posed" (2006, p. 6). The reason for this is closely linked to the next property.

Local interactions are typical of agent-based models. When agents interact with other agents, it is usually with their neighbors – those that are close to them in geographical space or in social space: their collaborators, colleagues, students, etc. The fact that not everyone interacts with everyone can make a significant difference in some situations. Simulations by Zollman (2007) and Grim (2009), for example, show important epistemic effects related to the non-universality of communication in scientific networks.

Bounded rationality, finally, states that: "Agents do not have global information, and they do not have infinite computational power. Typically, they make use of simple rules based on local information [. . .]" (2006, p. 6).

Scientists have sometimes been portrayed as somewhat irrational, uninformed, self-interested thinkers (e.g., Latour and Woolgar 1979; Hull 1988b). While this is slightly unpalatable to epistemologists who are concerned with perfect rationality, it has interesting consequences for models of science. Given agents that (like real scientists) have limited information and reasoning power, how can we set up the social structure of science for epistemic efficiency?

4.1.3 Some Methodological Considerations

Most of what applies to formal models in general (and that is covered elsewhere in this book) also applies to ABMs. In this section, we will focus on some issues that are specific to ABMs.

4.1.3.1 Micro vs. Macro

As we have hinted above, ABMs are concerned with the micro-level processes that give rise to observable, higher-level patterns. If an ABM can generate some macro-phenomenon of interest, then it can at least be considered a candidate explanation for it. When taken seriously, that possibility can become a requirement. This is what Epstein calls the generativist motto: "*If you didn't grow it, you didn't explain it*" (2006, p. 51). On this view, a pattern like Lotka's law (Lotka 1926) stands in need of explanation, and even an algebraic derivation of the law, like that of Herbert Simon (1955, p. 148), is still not sufficient for a complete explanation. One needs to supply the *mechanism* that generates the distribution. In the particular case of Lotka's law, that was achieved by Nigel Gilbert (1997), as we will see in Sect. 4.2.1.

Now this raises the question of what scale to choose for a model. The difference between micro and macro is relative to that choice. After all, if we were to grow a scientist from a collection of cells, the behavior of the scientist as a whole would be the macro-level. Now, it is assumed in *agent*-based modeling that the agent should be the micro-level, but what is an agent? Most models of science will focus on individual researchers as agents, but nothing prevents a modeler from focusing instead on research teams, labs, institutions or even whole countries. In Gilbert's model (oddly, perhaps) the papers themselves are the agents. In the end, it is left to the researcher to identify what Claudio Cioffi-Revilla, in a recent methodology paper, calls the "Cast of Principal Characters": "the main social entities themselves and their main interactions or causal dynamics" (2009, 30).

4.1.3.2 Details Matter

Once the target level has been chosen and the relevant entities identified, there remains the question of the amount of detail in which they must be modeled. The first ABMs usually had very simple agents. In Schelling's models, an agent's only properties were its position and its tolerance level. The interesting features of the model result from the relations and interactions between objects, not from the properties of the objects themselves. It is important to make sure, however, that such simplifications are not responsible for the behavior of the model.

To illustrate this caveat, we will use a model by Ryan Muldoon and Michael Weisberg (2010) looking at the distribution of cognitive labor over scientific projects. Given multiple projects, with different probabilities of success, there is an optimal assignment of scientists to projects: how can we ensure that the actual distribution of scientists approximate that optimum?

That question was previously studied by Philip Kitcher (1990) and Michael Strevens (2003) using analytical models. It is the purpose of Muldoon and Weisberg (2010) to show that some of the idealizing assumptions made by Kitcher and Strevens lead to results qualitatively different than if a more realistic model of the way the agents behave had been used. In other words, they fulfill only the first of these two requirements:

> [M]odels of cognitive labor must be simple enough for us to understand their dynamics, but faithful enough to reality that we can use them to analyze real scientific communities. (Muldoon and Weisberg 2010)

Kitcher and Strevens built their models using what Muldoon and Weisberg call the *marginal contribution/reward* (MCR) approach, in which each project is assigned a *success function*, "which represents the ability of the project to productively utilize the cognitive resources of scientists and turn those resources into the possibility of a successful outcome" (Muldoon and Weisberg 2010). Scientists working on a project that succeeds get a reward, according to a scheme that can be varied, so each scientist chooses to work on the particular project that maximizes his own expected reward. We are looking for the reward scheme that produces the best allocation.

Muldoon and Weisberg (2010) claim that Kitcher and Strevens' models rest on at least two unrealistic assumptions:

1. *Distribution assumption*: "every scientist knows the distribution of cognitive labor before she chooses what project to work on."
2. *Success function assumption*: each project's "success function, which takes as input units of cognitive labor (work from scientists) and outputs objective probabilities of success," is "known by all of the scientists in the model."

Both of these are assumptions of complete knowledge on the part of the scientists. To make their own model more realistic, Muldoon and Weisberg introduce complexifications in line with some of the characteristics we have seen in Sect. 4.1.2: agents do not have perfect knowledge (bounded rationality) and not every agent knows or believes the same things (heterogeneity).

Let us start with the distribution assumption. Muldoon and Weisberg's scientists are distributed on a grid (a torus, actually) of 35 × 35, where distance represents "communication distance." Scientists have a "radius of vision": they "see" the project choices of other agents within that radius. To mimic Kitcher and Strevens' perfect information scenarios, the radius of vision must be at least $\sqrt{578}$, the distance at which everyone sees everyone.[2] When Muldoon and Weisberg do that,

---

[2] For an agent standing exactly in the middle of a flattened 35×35 torus, euclidean distance to each corner $= \sqrt{17^2 + 17^2} = \sqrt{578} = 24.0416306$.

using Strevens' *Marge* reward scheme, where payoff is divided equally between all agents working on the successful project, they get the same results as Strevens:

> As the number of agents was increased, an incentive was created for a minority of scientists to work on the harder project. When the number of agents was increased further, scientists allocated themselves to both projects, and eventually the number of scientists working on the harder project overtook the number working on the easier project. (Muldoon and Weisberg 2010)

When Muldoon and Weisberg decrease vision, however – i.e., they relax the perfect knowledge distribution assumption – agents start to misallocate: when vision drops below seven, no one works on the harder project. From a collective point of view, this is far from optimal.

Now, for the success function assumption, Muldoon and Weisberg argue that it is very unrealistic that every scientist would know the objective probability of success of each project. Those probabilities should be subjective, and hence, vary from one scientist to another. In their model, Muldoon and Weisberg use a success probability function taken from Kitcher[3] which has an "easiness" parameter, and evaluation of that easiness is where agents differ. Muldoon and Weisberg assumed that the agents' beliefs about the easiness of a project follow a normal distribution where the mean is the objective probability of success of the project. A variance of zero in that distribution mimics the Kitcher/Strevens perfect information scenario and, again, a simulation with two projects (one easier, one harder) confirms that. But as soon as variance is introduced – i.e, as soon as some agents misjudge the probability of success – the resulting allocation is suboptimal.

Part of the appeal of models of science (and models in the social sciences at large, for that matter) is that once we have a good one, it can possibly be used to inform policy making. That is part of what Strevens is trying to do when he compares the *Marge* reward scheme (equal payoff for everyone on the successful project) to the *Priority* scheme that we use in reality (first successful scientist gets all the credit). In Strevens' model, *Priority* produces a better distribution of cognitive labor. In Muldoon and Weisberg's more realistic model, *Priority* does worse than *Marge*. The take home message is that it is important to get the details right. As the case of Muldoon and Weisberg show, and as we will further try to show in the next section, ABMs are a good way to do that.

4.2 What Has Been Done So Far?

We now move on to Gilbert's original model (1997), which is arguably the most well-known ABM of science. We will describe it in a fair amount of detail, and use it afterward to contrast other models.

---

[3] That is the logistic growth equation: $P = \frac{K}{1+e^{-rN}}$, where K is the maximum probability of success, N the number of scientists working on the project, and r the easiness of the project.

4.2.1 Gilbert's Original Model: Papers and Kenes

Gilbert's explananda are the quantitative regularities traditionally found in science. That includes Lotka's law, but also many features of "little science" pointed out by de Solla Price in *Little Science, Big Science* (1963): e.g., exponential growth of the number of papers and the fact that references in a paper tend to be to recently published literature.

Gilbert starts out with a simple model of a candidate mechanism for simulating Lotka's law (Lotka 1926). In Gilbert's words, Lotka's law states that "for scientists publishing in journals, the number of authors is inversely proportional to the square of the number of papers published by those authors" (1997, 4.1). Most authors publish only one or two papers, some of them publish a little more, and only a few publish more than 10. Herbert Simon (1955) describes the probability of a paper being published by a scientist already having i publications as $f(i) = a/i^k$ (where a and k are constants). Simon also found a constant probability $\alpha = n/p$ that the next article in a journal is by a previously unpublished author (where p is the number of papers in the journal and n is the number of authors[4]). Gilbert's proposed mechanism for generating that pattern is actually quite simple:

> 1. Select a random number from a uniform distribution from 0 to 1. If this number is less than α, give the publication a new (i.e., previously unpublished) author.
> 2. If the publication is not from a new author, select a paper randomly from those previously published and give the new publication the same author as the one so selected.
>
> (1997, 4.4)

Note that Gilbert does not actually make use of $f(i)$. If the publication is to be assigned to a previously published author, all authors have an equal chance of being selected. The data produced by Gilbert's model approximate Simon's estimates and actual bibliometric data very closely, even if the simulation is completely agnostic of the expected probability distribution of authors.

Note, also, that the model is centered on papers: they are, in a way, the "agents" in his simulation. That stays the case when Gilbert moves on to a more complex simulation, in which the papers actually have some sort of content. They each contain a "quantum of knowledge" that is represented by what Gilbert calls a "kene." A kene is basically a sequence of bits that could, in theory, be of any length. To allow display of kenes in a two-dimensional plane, however, Gilbert makes them 32 bits long, encoding two 16-bit integers for x, y coordinates on a 65,536 × 65,536 grid, allowing talk about the *location* of a kene or a paper (which is that of its kene).

"Kene" is chosen to sound like "gene," and the reason for that is that there is an "evolutionary" component in the process. At each time step, at least one paper reproduces itself, and other existing papers[5] also have a small constant probability

---

[4] There is actually a typo in Gilbert's paper, where he states that $\alpha = p/n$ (it should be the other way around).

[5] Though Gilbert does not mention it explicitly, the simulation has to be initialised with a certain number of seminal papers: e.g. 1,000.

$\omega = 0.0025$ of reproducing. The author of a new paper is either a new author (with probability α) or the author of the parent paper. The new paper initially has the same kene as its parent. The new paper also has references: it chooses, at random, other papers located within a radius of $\epsilon = 7{,}000$.

It is supposed that each reference has an influence on the original kene, such that the final kene of the new paper is a combination of the original kene and the kenes of the references. If you think about kenes as points in space, you can think of each of the references' kenes as having a gravitational field that "pulls" the kene of the new paper in its direction. More formally, given a random value m between 0 and 1, increasing monotonically with each reference:

$$x'_p = x_p + (x_r - x_p)\frac{1-m}{2} \quad \text{and} \quad y'_p = y_p + (y_r - y_p)\frac{1-m}{2}.$$

This more detailed model still produces the Lotka's law pattern for the distribution of papers per author, which is not surprising since the part of the mechanism that generates that distribution is almost the same.[6] The model also produces a highly skewed distribution of citations per author, and that also matches empirical data. The overall growth rate (driven by the probability ω of spawning a new paper) also fits de Solla Price's observations.

Finally, a new result of the more complex model is that we can now observe different clusters of papers in the space of possible kenes. This is a consequence of the evolutionary mechanism chosen by Gilbert, where each new paper falls in the vicinity of his parent. Those clusters are interpreted by Gilbert as representing different specialities in a field. A problem with that interpretation is that the position of the kene is not taken into account when the paper "chooses" its author. It would be fairly straightforward, however, to take that factor into account (for example, by having the probability of a particular author being selected increase if he has recently written a paper in the area of the new paper.) (Figs. 4.6 and 4.7)

4.2.2 Follow-Ups and Other Models

While fairly simple, Gilbert's model is a striking example of the possibilities of agent-based modeling of science. Gilbert himself, with collaborators Andreas Pyka and Petra Ahrweiler, took the idea further in a series of papers on innovation networks (Gilbert et al. 2001, 2007; Pyka et al. 2002, 2007, 2009; Ahrweiler et al. 2004). Börner et al. (2004) also have a model called TARL (for "topics, aging, and recursive linking") where they dynamically generate a network of co-authorship relations in addition to a citation network similar to that of Gilbert, and which they validated against a PNAS data set of significant size. Gilbert's model also

---

[6] The only difference is that authors now "retire" after a random number of time steps (where the maximum is $\phi = 480$).

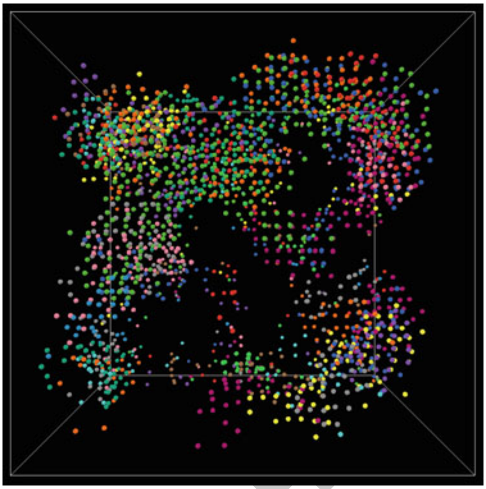


Fig. 4.6 The 2D landscape of papers in Gilbert's simulation, showing the clustering into "specialities"

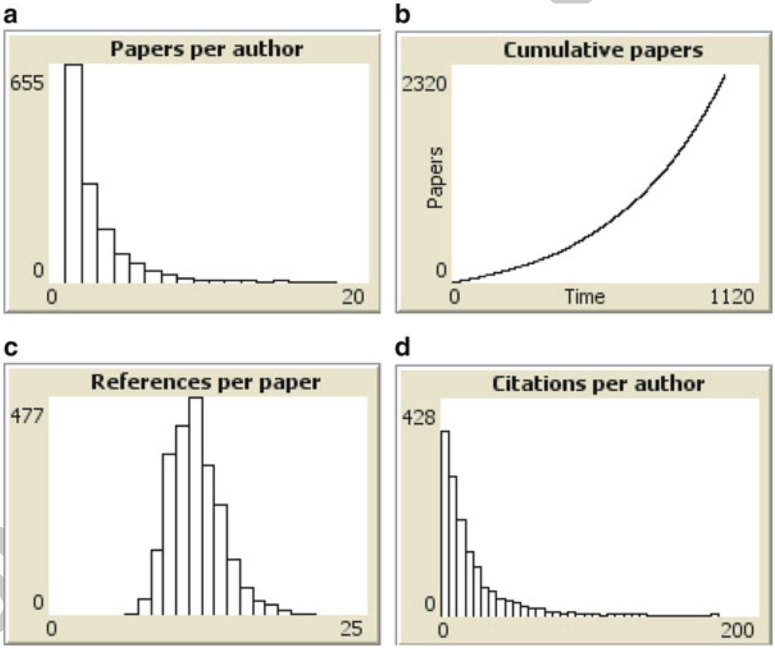


Fig. 4.7 Numerical relationships in a sample run of Gilbert's simulation. Figure 4.7 shows the approximation of Lotka's law. The figures are taken from Gilbert's NetLogo replication of his original model, freely available at http://www.openabm.org/model-archive/ssas

directly inspired models in fairly different areas (e.g., Boudourides and Antypas 2002). In this section, we will look at models by Sun and Naveh (2009) and Edmonds (2007). We will then move on to models by Weisberg and Muldoon (2009) and Grim (2009) that, while not concerned with the specific bibliometric patterns explored by Gilbert, are closely related to an idea he almost touched on with his spatial distribution of kenes: that of an epistemic landscape.

4.2.2.1 A View from Cognitive Science

The fact that scientists play only a very small role in Gilbert's model can be a target for criticism. It is hard to accept the idea that the only difference between the author who published only one paper and the one who published 15 is that the latter got lucky in that more papers selected her.

Cognitive scientists Sun and Naveh (2009), in particular, have been critical: "Gilbert's model lacks agents capable of meaningful autonomous action" (2007, p. 142). They have attempted to provide a more realistic model, where "authors are not merely passive placeholders, but cognitively capable individuals whose success or failure depends on their ability to learn in the scientific world" (2006, p. 321). In order to achieve that, they use a cognitive architecture they call CLARION, an acronym that stands for "Connectionist Learning with Adaptive Rule Induction ON-line." The full name is actually a fairly good description of what CLARION does. It is a *hybrid* architecture: it has a learning mechanism implemented in an artificial neural network, but it can extract explicit symbolic rules from what it has learned at the connectionist level and use these rules to drive its behavior. We will not go into the details of CLARION (see Sun (2006) for an overview and Sun (2003) for a detailed description), but it is meant to be cognitively realistic. Sun himself has argued extensively for such hybrid systems (Sun 2002), and what has come to be called "dual process theories" are increasingly prevalent in cognitive science (Evans 2008) (Fig. 4.8).

In Sun and Naveh's model, as expected, it is now each scientist that selects an idea to replicate, and not the other way around. The scientists also select the neighboring ideas that they use to modify the original idea, but they do not stop at that: they also *optimize* the resulting idea on their own, by searching the space around it for slightly better positions. (We are still talking about ideas as vectors in a multidimensional space, just like Gilbert's kenes.) The fact that such an optimization is going on implies that, as opposed to what we had in Gilbert's model, some ideas are better than others. Sun and Naveh name a few properties over which ideas differ: clarity, insightfulness, empirical evidence, theoretical results and application potential. Agents all have "subjective functions" for these different properties of ideas: functions that they refine throughout the simulation, trying to approximate the "communal" functions that determine if a paper gets published. No agent has direct access to the communal functions: all they have is the feedback they get from the submission of a paper: i.e., whether it is accepted or not. They use this feedback to optimize two tasks: (1) choosing the focal idea and (2) choosing the pull ideas. Agents that fail to publish enough are removed from the simulation and replaced by new agents. In their model, it is that learning process, instead of luck,

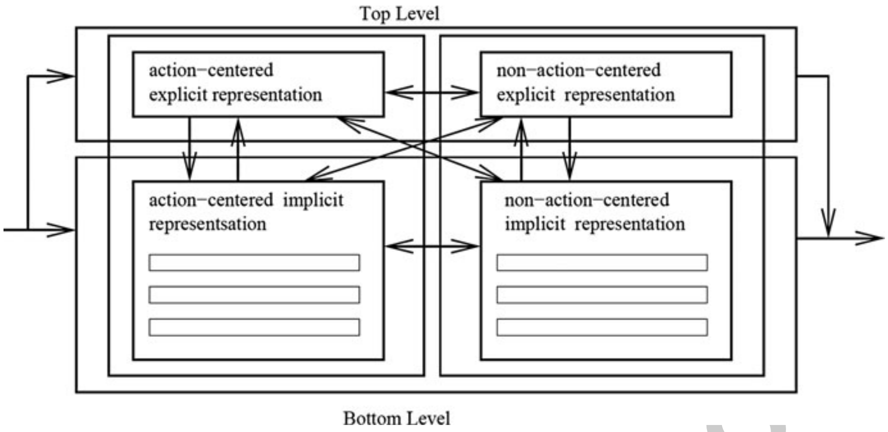


Fig. 4.8 Sun and Naveh's CLARION architecture, showing the interaction between explicit (symbolic) representations and implicit (connectionist) representations

Table 4.2 Number of authors contributing to *Chemical Abstracts*

# of Papers	Actual	Simon's estimate	Gilbert's simulation	CLARION simulation
1	3991	4050	4066	3803
2	1059	1160	1175	1228
3	493	522	526	637
4	287	288	302	436
5	184	179	176	245
6	131	120	122	200
7	113	86	93	154
8	85	64	63	163
9	64	49	50	55
10	65	38	45	18
11 or more	419	335	273	145

that is responsible for the difference in the number of publications by each agent AQ3 (Tables 4.2 and 4.3).

Sun and Naveh's results also match the empirical data, but not as closely as Gilbert's model. There is, however, a good reason for that:

> We put more distance between mechanisms and outcomes, which makes it harder to obtain a match with the human data. Thus, the fact that we were able to match the human data shows the power of our cognitive agent-based approach compared to traditional methods of simulation. (Naveh and Sun 2007, p. 200–201)

Sun and Naveh's model allows them to study the effect of cognitive differences on the success of the whole community. The latter is measured by the total number of papers published. In Gilbert's model, that number was a direct result of the parameter ω (the probability that a paper would spawn a new paper). Here, it is a

Table 4.3 Number of authors contributing to *Econometrica*

# of Papers	Actual	Simon's estimate	Gilbert's simulation	CLARION simulation
1	436	453	458	418
2	107	119	120	135
3	61	51	51	70
4	40	27	27	48
5	14	16	17	27
6	23	11	9	22
7	6	7	7	17
8	11	5	6	18
9	1	4	4	6
10	0	3	2	2
11 or more	22	25	18	16

result of the ability of the agents to learn the communal rules of publication. Those cognitive parameters are many (e.g., the learning rate of the agents, the probability of using implicit vs. explicit learning, the randomness of the local search process), and they all have significant effects on the overall number of papers.

4.2.2.2 Science as a Distributed Cognitive System

Cognition can also be conceived as going beyond the individual level. Some philosophers (e.g., Thagard 1993b; Giere 2002; Cummins et al. 2004; Magnus 2007) have been claiming that science as a whole should be thought of as a distributed cognitive system. Joshua Epstein goes even further and claims that:

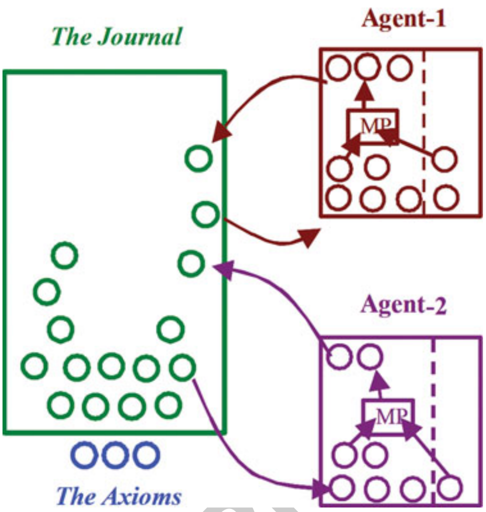
> The agent-based approach invites the interpretation of society as a distributed computational device, and in turn the interpretation of social dynamics as a type of computation. (Epstein 2006, p. 4)

Bruce Edmonds (2007) takes that idea seriously. He proposes an agent-based model of science as a distributed theorem prover. In contrast to what we have seen so far, the knowledge acquired by Edmond's agents is something highly structured: true sentences in a formal system, namely, propositional logic. In effect, agents are trying to come up with new theorems by combining existing items of knowledge (premises) into new ones by inference. In Edmond's model, agents are confined to using the *modus ponens*[7] inference rule: i.e., $((p \rightarrow q) \wedge p) \vdash q$. Every agent has a store of knowledge – sentences that can be used as premises for new inferences. Those sentences come from inferences made by the agent, but also from a public repository of knowledge: a "journal," in which agents publish some of the theorems they find. At each time step, every agent:

---

[7] In (almost) plain English, the *modus ponens* rule says that if you know some proposition p to be true and you also know that *if p, then q*, you are allowed to deduce that q is true.

Fig. 4.9 A representation of Edmond's agents interacting with the knowledge store



this figure will be printed in b/w

1. Replaces some of the sentences in its private store by sentences from the journal.
2. Tries to combine sentences from its private store and adds the result of successful inferences to its private store.
3. Submits previously unpublished items from its private store to the journal.

At the end of a time step, the journal ranks the received submissions as a weighted sum of "the extent to which a formula had the effect of shortening formula when applied as the major premise in MP; the shortness of the formula itself; the past publishing success of the author; and the fewness of the number of distinct variables in the formula" (Edmonds 2007).

The success of the community is evaluated according to the number of useful theorems it can find in a given number of time steps. "Useful," here, means *really* useful: the system is judged against a list of 110 target theorems taken from logic textbooks (Fig. 4.9).

One of the interesting findings of Edmond is that the number of useful theorems found is fairly independent from the publication rate of the journal (i.e., the number of submission it accepts each turn). Another interesting finding is the disparity between individual agents: some of them publish a lot more than others. While not quite as "Lotka-like" as Gilbert's or Sun and Naveh's results, it is still a fairly skewed distribution.

Notwithstanding the detailed dynamics of Edmond's model, an important insight is that ABMs of science can be made to work on "real world" science problems. Of course, propositional calculus (especially the "one inference rule version" used by Edmond) is somewhat of a toy problem, but we can imagine a system working on more complicated, more realistic problems. These would have to be well-defined formal problems as opposed to the open-ended research that scientists are usually involved in. The idea is not to use ABMs of science to computationally solve new problems – that would be the job of computer scientists,[8] not social science modelers. Nonetheless, it seems conceivable that, if we pay close attention to the analogues between computational problem-solving algorithms and the scientific process, ideas from one could be used to improve the other, and vice versa.

4.2.2.3 Science as an Epistemic Landscape

Most of the models mentioned assume that there is some sort of (possibly highly dimensional) space that the agents are trying to explore. Some positions in this space are considered better than others and agents are trying to find these positions. Different models assume different semantics for the space, the most common being that proximity in the space corresponds to some sort of conceptual, theoretical, or pragmatical similarity.

For that space to be interpreted as a *landscape*, however, one-dimension must stand for "height." Kenes in Gilbert's model were situated on a two-dimensional space, but no value was attached to them; the space was flat. In Sun and Naveh's model, ideas were also situated on a two-dimensional plane, but different ideas had different values: some were clearer, better supported empirically, etc. If you collapse all of these values in a single weighted sum, you get a third-dimension: the height of the landscape. Of course, you can also have a n-dimensional space, as long as there is one-dimension that you are trying to maximize.[9] What Sun and Naveh did not insist on, however, is how the shape of that landscape affects the dynamics of science. To illustrate some of these dynamics, we will look at another model by Weisberg and Muldoon (2009), one that builds on the work presented in Sect. 4.1.3.2 (Fig. 4.10).

Agents in Muldoon and Weisberg (2010) were[10] situated in space, but distance in that space represented communication distance between researchers, not the value of the projects they were working on. Weisberg and Muldoon are still interested in the division of cognitive labor, but this time, instead of looking at just two projects with different probabilities of success, they look at the whole range of different approaches available to scientists within a research topic. As you might have guessed, these approaches are represented by the position of a scientist agent

---

[8] It has already been shown that some A.I. programs are capable of scientific reasoning. The classic example would be BACON (Langley et al. 1981), which "rediscovered" Snail's law of refraction, conservation of momentum, Black's specific heat law, and Joule's formulation of conservation of energy. The PI program (Thagard and Holyoak 1985) achieves similar results, but is perhaps more suited to an agent-based approach (Thagard 1993a, ch. 10).

[9] You could also have many dimensions that you are trying to optimize. Those problems are known as "multiobjective optimization problems" (Steuer 1986; Sawaragi et al. 1985) In those cases, you are looking for the "pareto front": the set of positions in space that are not "strictly dominated" by any other. We will leave those complexities aside.

[10] Weisberg and Muldoon (2009) was published before Muldoon and Weisberg (2010), but the latter reports on an earlier model.

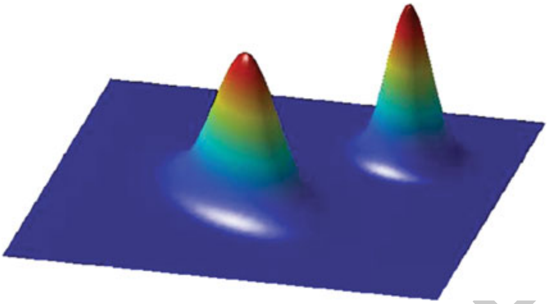


Fig. 4.10 The epistemic landscape used in (Weisberg and Muldoon 2009). The vertical axis represents "epistemic significance"

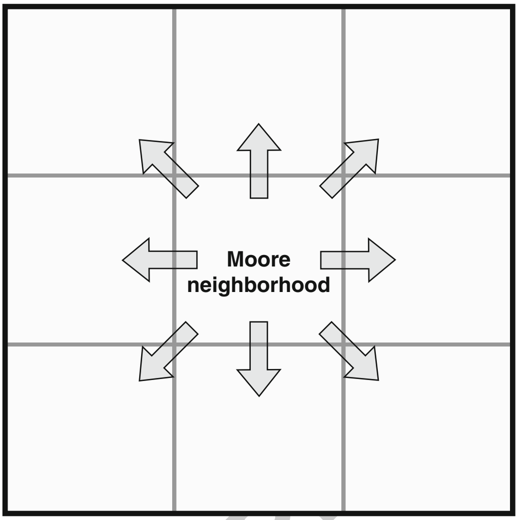
in two-dimensional space.[11] The third-dimension is what they call the "epistemic significance" of the approaches. The goal of the agents is, of course, to find the highest peaks of significance in the landscape. The landscape used by Weisberg and Muldoon has two peaks, generated by two Gaussian functions. The way agents move around the landscape depends on the strategies (i.e., the rules of behavior) they adopt. Investigating the way populations with different mixes of strategies explore the landscape is the authors' purpose. They look at three different strategies. Here they are, in very general terms:

- **Controls** are basically "hill climbers": they set a direction and move forward as long as they get better results. If they get worse results, they backtrack and set a new, random, direction. They are "controls" in the sense that they do not take into account what the other agents on the landscape are doing, and the authors are mostly interested in the dynamics introduced by interactions between agents.
- **Followers** start by looking for all the squares in their Moore neighborhood (see Fig. 4.11) that have previously been visited and have a greater significance than their current approach. If there are such squares, they will move to the best among those (breaking ties at random). If there are none, they will look for unvisited squares and choose one at random. (In other words: they will only innovate if they *have* to.) Finally, if all the neighborhood squares have already been visited and none is better than their current one, they stop.
- **Mavericks** are a little bit like controls in that if their current location is worse than their previous one, they will backtrack and change direction. But if their new approach is equal or better to the previous one, they will move to an *unvisited* spot

---

[11] Weisberg and Muldoon, like Gilbert before them and, as we will see in Sect. 4.2.2.4, Patrick Grim, leave the exploration of higher-dimensional space for "further research." The main advantage of 3D landscapes is, of course, that they can be visualised easily. They also simplify programming and keep computations light. It would be interesting, nonetheless, to see a detailed study of the impact of high-dimensionality on some models. The "curse of dimensionality" is a problem for many optimization tasks, and computer scientists are developing special algorithms and techniques to deal with it (e.g., Powell 2007), so it is conceivable that it would make a difference in the results of the simulations we are looking at. Muldoon and Weisberg have both (independently) been tackling that issue, but have not published about it yet.

Fig. 4.11 A Moore neighborhood simply consists of the nine squares surrounding a particular position on a grid



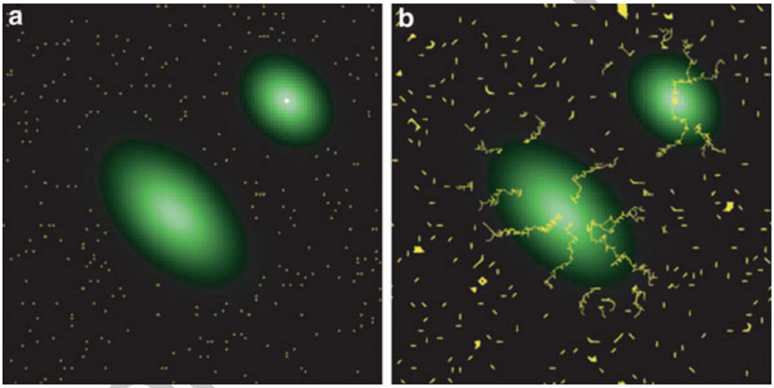


Fig. 4.12 Initial (**a**) and later (**b**) position of followers on the epistemic landscape during a simulation. The trails indicate the paths followed by agents

in their Moore neighborhood (choosing one at random if there are many). Only if there is no unvisited spot will they act like followers and choose the best-known approach around them.

Controls by themselves are not very efficient. They eventually find the peaks, but since they cannot learn from one another, it takes a lot of time steps before they get there. Followers alone do even worse: they get stuck in low-significance areas pretty quickly. Unless they are lucky, they will just follow each other around. Mavericks, on the other hand, are very efficient: they always find the peaks, and they find them a lot faster than controls (Figs. 4.12–4.14).

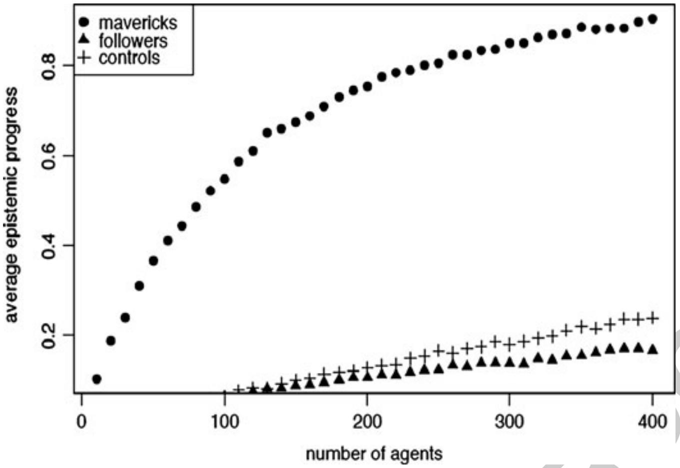


Fig. 4.13 Number of agents *vs* epistemic progress in homogeneous populations

Things get more interesting when you start looking at mixed populations. Even adding a single maverick in a population of followers makes a significant difference: mavericks help the followers get unstuck. The more mavericks you add, the more performance improves, until you reach 100% mavericks, which is the optimum. In the real world, though, a balance between followers and mavericks is probably needed. Followers seem well-suited to what Kuhn (1962) called "puzzle solving": finding solutions to very specific problems with well-defined methods. Being a maverick is probably more risky for the individual: wandering off the beaten path and possibly failing can be very costly for one's career.

Weisberg and Muldoon's work show that the way researchers deal with the results of other agents around them makes a difference for the overall success of the community.

4.2.2.4 Epistemic Networks

Kevin Zollman has done some pioneering work on simulating the effect of the social structure of the scientific community on its epistemic performance. The matter is important, he says, because:

> Once one fully articulates a theory of individual epistemic rationality, it is still an open question what the optimal community structure is for these agents – the individualistic question is only part of the answer. (Zollman 2007)

To try to answer the question of "optimal community structure," Zollman simulates the behavior of networks of scientists trying to choose between two distinct hypotheses, given limited information. Individual scientists can only communicate their results to their immediate neighbors. We will not go into detail about his

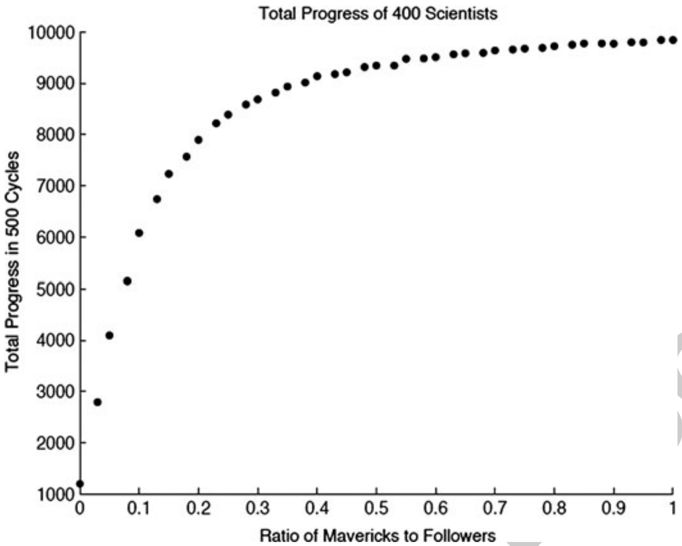


Fig. 4.14 Ratio of mavericks and followers *vs* epistemic progress in mixed populations

experiment, but what he found was essentially that a more connected network will converge much more rapidly on an hypothesis, but is much more likely to converge on the *wrong* hypothesis: there is an important trade-off between speed and accuracy.

Partly inspired by Zollman, Patrick Grim also did some work about how the social structure of science affects its results:

> How does an individual figure out the structure of the world? The truth is that no individual does. It is cultures and communities that plumb the structure of reality; individuals figure out the structure of the world only as they participate in the epistemic networks in which they are embedded. (Grim 2009)

The main difference with Zollman is that Grim's agents, much like Weisberg and Muldoon's, are looking for the best hypothesis on an *epistemic landscape*. But, instead of seeing the results obtained by other agents around them on that landscape (like Weisberg and Muldoon's agents), Grim's agents see the results of those with whom they are connected in a social network (like Zollman's agents). At each time step, an agent has a 50% probability of modifying their current hypothesis by moving it halfway towards the best hypothesis amongst those of their connections.

That allows Grim to test for the best network structure amongst many that are prevalent in the social network literature: ring, small world, wheel, random, and complete networks.[12] What he finds is that the ring network performs the best, while the complete network performs the worst. In general (and he shows this with random networks), above a very low threshold, adding links to a network decreases performance. That is consistent with Zollman's results.

Analyzing his results, Grim speculates that for at least some problems, the scientific network of the seventeenth century, where communications between researchers were few and far between, might have been better adapted than the fully connected, round the clock, social network of twenty-first century science.

In fact, what happens in the fully connected networks is similar to what happens with Weisberg and Muldoon's followers: researchers stay confined to regions of the landscape that are already explored. That makes the community vulnerable to getting stuck on peaks of non-optimal epistemic value (like the one on the left of Fig. 4.15) because everyone will converge on the best initial hypothesis, and no one will explore further once they have reached it. What Grim needs, it seems, is a few mavericks: researchers who will deliberately avoid duplicating their peers' hypotheses (Fig. 4.16).

4.3 Where Should We Go From Here?

Having taken a look at a very diverse (but maybe not fully exhaustive) list of agent-based models of science, we will end this chapter by trying to identify a few questions that might benefit from agent-based modeling and, finally, point out a few methodological issues faced by modelers today.

4.3.1 Directions for Future Research

While the details of the process are generally not agreed upon, many thinkers concur that science, somehow, evolves (Popper 1959; Toulmin 1972; Campbell 1974; Hull 1988b). This is an important idea, as far as ABMs go, because they are especially well suited for evolutionary models. The evolutionary notion of fitness landscapes is closely related to the notion of epistemic landscapes that we have seen in Sect. 4.2.2.3. Besides, evolutionary ABMs of science can draw heavily from the field of genetic algorithms and related techniques (Holland 1975; Luke 2010).

Though Gilbert's simulation has a small evolutionary component, in that each kene is descended from a parent kene, an important element is missing: which paper gets to "reproduce" is not a function of the content of the kene (papers are just randomly selected for reproduction). If, on the other hand, you had differential reproduction, based on the position of the kene in an epistemic landscape similar to

---

[12] Animations of the networks should be viewable on Grim's website, at: http://www.pgrim.org/ABMScience.

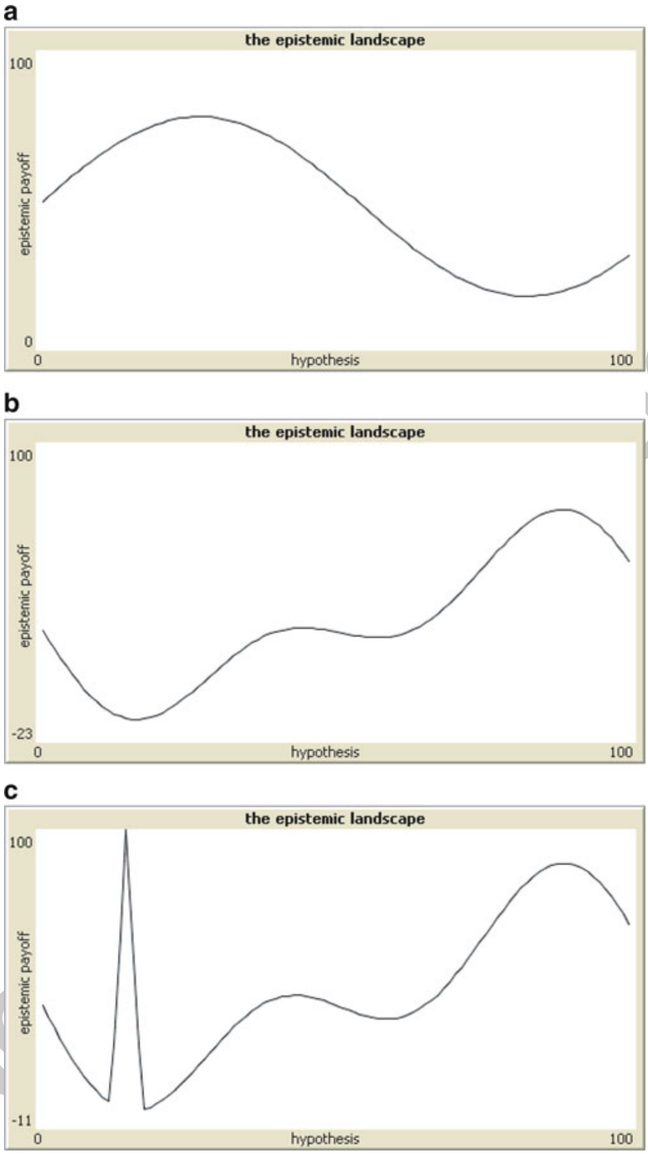
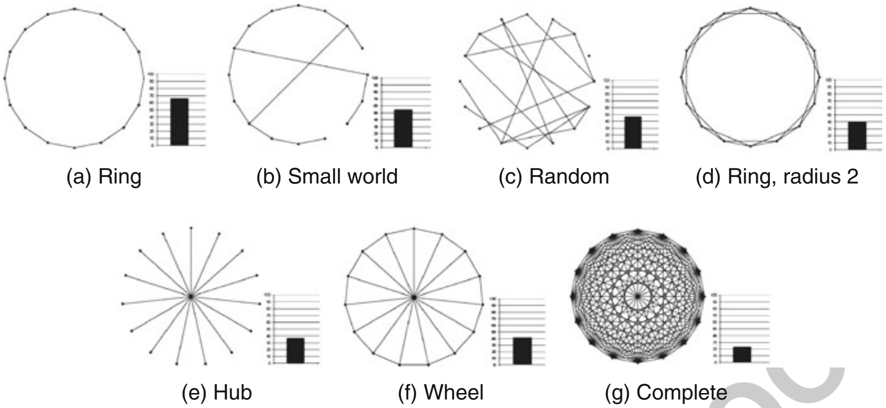


Fig. 4.15 Different shapes of 2D epistemic landscapes in the Grim simulation. Notice how landscape Fig. 4.15 is deceiving for the agents, and as such, considerably harder than the other two

those used by Weisberg and Muldoon of Grim, then you would get *adaptation*: i.e., the kenes (or papers, ideas, theories, etc.) with a higher position on the landscape would tend to out-reproduce the others. To our knowledge, this idea has not been fully explored yet.



(a) Ring (b) Small world (c) Random (d) Ring, radius 2

(e) Hub (f) Wheel (g) Complete

Fig. 4.16 Percentage of runs where the community finds the optimal hypothesis for different network structures on the landscape Fig. 4.15. In general, the less connected the network, the better its performance

Another potentially important evolutionary component in science is the teaching process: a supervisor transmits his ideas to his students, which in turn (if they are successful) will teach a subsequent generation of students, and so on. You can build a genealogy of researchers just like you can build family trees. This can bring interesting insights – for example, the idea that Hull (1988a) calls "conceptual inclusive fitness," which is based on Hamilton's biological notion of "inclusive fitness" (1964): just as altruistic behavior towards one's relatives promotes the replication of shared genes, altruistic behavior towards one's graduate student (co-authoring papers with her, sending her to conferences, helping her find a good academic position) promotes the replication of ideas transmitted to her. Again, this is something that an agent-based model would be well suited to explore.

Another interesting idea from David Hull is the trade-off between "credit" and "support." The main premise behind this one is that scientists' primary motivation is getting credit for their theories: that is, mostly, recognition from their peers. (That credit can then be "cashed-in" in various ways.) Credit is mainly attributed by way of citation: if you base part of your work on somebody else's work, then you share the credit by citing his work. But as Hull puts it:

> One cannot gain support from a particular work unless one cites it, and this citation automatically both confers worth on the work cited and detracts from one's own originality. Scientists would like total credit and massive support, but they cannot have both. Science is so organized that scientists are forced to trade off credit for support. (2001, p. 100–101)

There is a whole continuum of strategies, between high risk/high reward and low risk/low reward, which can be adopted by individual scientists. The distribution of these strategies within a population of researchers should have an impact on the scientific enterprise as whole, and what kind of impact is something that would be interesting to assess using ABMs. One could also take into accounts feedback loops related to these phenomena: e.g., the more credit you already have, the more likely you are to be cited, and thus, gain even more credit.[13]

A final suggestion as to what should be addressed by ABMs of science is the *dynamics* of research networks. We have seen, with Grim (2009) and Zollman (2007), that network structure has important effects on social epistemic processes, but how do these networks form? How do they change over time? Does having a lot of credit allow a researcher to attract good collaborators, which would, in turn, provide him with even more credit? Can we simulate the formation of invisible colleges? What about the rivalry between different communities of researchers? All these, and many more questions, could potentially be studied using ABMs.

The questions raised in this section are just a (fairly arbitrary) sample of what could potentially be done by using agent-based models of the scientific process, but going forward, there are also methodological issues to be addressed.

4.3.2 Methodological Issues

Agent-based modeling in the social sciences is still a fairly immature field, and ABMs of science even more so. Many researchers are writing about methodological issues (e.g. Axtell et al. 1996; Cioffi-Revilla 2009; Epstein and Axtell 1996; Gilbert and Troitzsch 2005), but a common methodological framework for model building has yet to emerge. In the meantime, many concerns come to mind.

Most of the models we have seen in this chapter have overlapping but slightly different features. We have compared them to one another, but from a very high-level, qualitative point of view. There is no doubt that the field would benefit from more systematic comparisons between models (see Axtell et al. 1996). Independent replication of existing models is also a useful – but seldom undertaken – endeavor, which can reveal incoherence (or at least ambiguity) in the original description of a model.

One can also ask if it is time to try to integrate all of these models into a single framework (maybe open source?) that everyone can thoroughly explore and even extend? (In other words, should we continue to be mavericks, or are we ripe for some followers?)

Agent-based modeling of science calls for knowledge from many different disciplines: scientometrics, information science, economics, game theory, artificial intelligence, social network analysis, evolutionary computation, cognitive science in general and even cognitive anthropology, all have something to contribute. This probably requires the assembling of interdisciplinary teams and that is a challenge in itself.

---

[13] The author of the present chapter is currently working on a model trying to take these issues into account as part of a PhD thesis entitled: "Simulating Science: an Agent-Based Model of Scientific Evolution". (Université du Québec à Montréal, Département de philosophie).

Also, though we did not raise the issue in the previous sections, the fact is that agent-based simulations are computer programs, and building one is by no means trivial. There are many tools that one can use to build an ABM: it can be built from scratch using any programming language, or it can use a powerful low-level library like MASON (Luke et al. 2005)[14] or a high-level framework like NetLogo.[15] Other common frameworks are Repast[16] and Swarm,[17] but you will find many others, with different degrees of simplicity, generality and popularity, in Nikolai and Madey (2009). Lots of questions may be asked: Is this multiplicity of tools a good or a bad thing? How does it affect collaboration between modelers? How does it affect reproducibility of the results? Can the models be fully described in abstract, mathematical language, or does implementation matter? Would we be better off with a single framework (maybe targeted specifically for science modeling)?

Finally, agent-based modeling of science needs to find a place for itself amongst traditional mathematical models and scientometrics. Just as it does for traditional models, scientometrics provides explananda for ABMs. ABMs are able to generate massive amount of data that can then be analyzed and visualized using the best available tools from scientometrics. To our knowledge, this has not fully been done yet, though Börner et al. (2004) took a significant step in the right direction. But still, it is an area where ABMs of science are sorely lacking.[18]

As for the relationship between ABMs and traditional analytical models, we have seen in Sect. 4.1.3.2 that ABMs can be used to challenge some idealizations made by other models. Hopefully, this can lead to a process of back-and-forth exchange that will be profitable for both types of models.

Key points

In their current state, agent-based models of science do not provide all that much in the way of direct policy recommendations. Nonetheless, some of the models we have seen point towards a few key insights that need to be recognized:

- In all knowledge-seeking systems, there is a trade-off between exploitation and exploration: a delicate balance between fine-tuning the knowledge you already have and striving for completely new knowledge. As Weisberg and Muldoon (2009) have shown, a population of scientists needs at least a few

---

[14] http://cs.gmu.edu/~eclab/projects/mason/

[15] http://ccl.northwestern.edu/netlogo/

[16] http://repast.sourceforge.net/

[17] http://www.swarm.org/

[18] To be fair, Gilbert et al. (2007) and Sun and Naveh (2009) do compare their results to scientometric data, but it is a *very* small dataset.

"mavericks," and that should be taken into account with things like funding decisions.

- Closely related to that first point is the issue of the division of cognitive labor: we want scientific resources to be allocated to different projects in a way that is optimal for the community as a whole. Individual incentives are a useful tool to try to achieve that, but as Muldoon and Weisberg (2010) have shown, some of them might not be as efficient as we think they are. The long reigning *Priority Rule*, for instance, might be due for a reevaluation.
- Scientists are part of communities, and the structure of these communities matter. The results we have so far regarding this question tend to show that too much communication between scientists might lead to premature agreement on some issues (Grim 2009). If that is indeed the case, the pressure to publish early and often may be having adverse effects on the performance of the science system.
- The concept of an "epistemic landscape" is probably a new one for most policy-makers, but it has far reaching implications: different policies are likely to have different effects on different epistemic landscapes, so the shape of the landscape should be taken into account when trying to influence the science system. It is not clear yet how to map the shape of the landscape for any particular domain, but this is a question that is likely to be at the forefront of "science of science" research in the coming years.

Those various insights show at least the potential of agent-based models of science, so one last recommendation should be:

- Agent-based models should become part of the policy-maker's toolbox, as they enable us to capture a kind of complexity that is not easily tackled using analytical models. While they are still in their infancy, they open up a new range of possibilities for investigating the science system.

References

Ahrweiler P, Pyka A, Gilbert N (2004) Simulating knowledge dynamics in innovation networks (SKIN). Volkswirtschaftliche Diskussionsreihe, no. 267. Universität Augsburg, Institut für Volkswirtschaftslehre, Augsburg, available online at the URL: http://www.wiwi.uni-augsburg.de/vwl/institut/paper/267.htm

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211(4489):1390–1396 (DOI: 10.1126/science.7466396)

Axtell R, Axelrod R, Epstein JM, Cohen MD (1996) Aligning simulation models: A case study and results. Comput Mathematical Organization Theory 1(2):123–141 (DOI:10.1007/BF01299065)

Bigbee A, Cioffi-Revilla C, Luke S (2007) Replication of sugarscape using MASON. In: Terano T, Kita H, Takahashi S, Deguchi H (eds) Agent-based approaches in economic and social complex systems IV: Post proceedings of the AESCS international workshop 2005. Agent-Based Social Systems, vol 3. Springer, Tokyo, pp 183–190 (DOI:10.1007/978-4-431-71307-4_20)

Börner K, Maru JT, Goldstone RL (2004) The simultaneous evolution of author and paper networks. Proc Natl Acad Sci USA 101(Suppl 1):5266–5273 (DOI:10.1073/pnas.0307625100), also available as arXiv preprint arXiv:cond-mat/0311459. The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci USA 101(Suppl 1):5183–5310

Boudourides M, Antypas G (2002) A simulation of the structure of the World-Wide Web. Sociol Res Online 7(1), available online at the URL: http://www.socresonline.org.uk/7/1/boudourides.html

Campbell DT (1974) Evolutionary epistemology. In: Schilpp PA (ed) The philosophy of Karl Popper. The Library of Living Philosophers, vol 14. Open Court Publishing, La Salle, IL

Cioffi-Revilla C (2009) Simplicity and reality in computational modeling of politics. Comput Mathematical Organization Theory 15(1):26–46 (DOI:~10.1007/s10588-008-9042-2)

Cristelli M, Pietronero L, Zaccaria A (2011) Critical overview of agent-based models for economics. ArXiv preprint arXiv:1101.1847

Cummins R, Poirier P, Roth M (2004) Epistemological strata and the rules of right reason. Synthese 141(3):287–331 (DOI:B:SYNT.0000044992.91717.aa, stable JSTOR URL: http://www.jstor.org/stable/20118485)

Edmonds B (2007) Artificial science: A simulation to study the social processes of science. In: Edmonds B, Troitzsch KG, Iglesias CH (eds) Social simulation: Technologies, advances and new discoveries. IGI Global, Hersey, PA, Chap 5, pp 61–67 (DOI:10.4018/978-1-59904-522-1.ch005), also available online at the URL: http://cogprints.org/4263

Epstein JM (2006) Generative social science: Studies in agent-based computational modeling. Princeton Studies in Complexity. Princeton University Press, Princeton, NJ

Epstein JM, Axtell R (1996) Growing artificial societies: Social science from the bottom up. Complex Adaptive Systems. Brookings Institution Press, Washington, DC; MIT Press, Cambridge, MA

Evans JSBT (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 59(1):255–278 (DOI:10.1146/annurev.psych.59.103006.093629)

Gardner M (1970) Mathematical games: The fantastic combinations of John Conway's new solitaire game 'Life'. Sci Am 223(4):120–123 (DOI:10.1038/scientificamerican1070-120)

Giere RN (2002) Scientific cognition as distributed cognition. In: Carruthers P, Stich S, Siegal M (eds) The cognitive basis of science. Cambridge University Press, Cambridge, pp 285–299 (DOI:10.1017/CBO9780511613517.016)

Gilbert GN, Troitzsch KG (2005) Simulation for the social scientist, 2nd edn. Open University Press, Maidenhead

Gilbert N (1997) A simulation of the structure of academic science. Sociol Res Online 2(2), available online at the URL: http://www.socresonline.org.uk/2/2/3.html

Gilbert N, Pyka A, Ahrweiler P (2001) Innovation networks – a simulation approach. J Artif Societ Soc Simul 4(3):8, available online at the URL: http://jasss.soc.surrey.ac.uk/4/3/8.html

Gilbert N, Ahrweiler P, Pyka A (2007) Learning in innovation networks: Some simulation experiments. Physica A 378(1):100–109 (DOI:10.1016/j.physa.2006.11.050)

Grim P (2009) Threshold phenomena in epistemic networks. In: Hadzikadic M, Carmichael T (eds) Complex adaptive systems and the threshold effect: Views from the natural and social sciences: Papers from the AAAI Fall Symposium, November 5–7, 2009, Arlington, Virginia. AAAI Technical Report FS-09-03. The AAAI Press, Menlo Park, CA, pp 53–60, available online at the URL: http://aaai.org/ocs/index.php/FSS/FSS09/paper/download/916/1234

Hamilton WD (1964) The genetical evolution of social behaviour. J Theor Biol 7(1):1–52 (DOI:10.1016/0022-5193(64)90038-4)

Holland JH (1975) Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence. University of Michigan Press, Ann Arbor, MI

Hull DL (1988a) A mechanism and its metaphysics: An evolutionary account of the social and conceptual development of science. Biol Philos 3(2):123–155 (DOI:10.1007/BF00140989)

Hull DL (1988b) Science as a process: An evolutionary account of the social and conceptual development of science. Science and Its Conceptual Foundations. University of Chicago Press, Chicgo, IL

Hull DL (2001) Science and selection: Essays on biological evolution and the philosophy of science. Cambridge Studies in Philosophy and Biology. Cambridge University Press, Cambridge

Kitcher P (1990) The division of cognitive labor. J Philos 87(1):5–22 (stable JSTOR URL: http://www.jstor.org/stable/2026796)

Kuhn TS (1962) The structure of scientific revolutions. University of Chicago Press, Chicago, IL

Langley P, Bradshaw GL, Simon HA (1981) BACON.5: the discovery of conservation laws. In: Drinan A (ed) Proceedings of the seventh international joint conference on artificial intelligence: IJCAI-81, 24–28 August 1981, University of British Columbia, Vancouver, BC, Canada. Morgan-Kaufmann Publishers, San Mateo, CA, vol 1, pp 121–126, available online at the URL: http://ijcai.org/PastProceedings/IJCAI-81-VOL1/PDF/025.pdf

Latour B, Woolgar S (1979) Laboratory life: The social construction of scientific facts. Sage Library of Social Research, vol 80. Sage Publications, Beverly Hills, CA

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12): 317–323

Luke S (2010) Essentials of metaheuristics. Lulu, available online for free at the URL: http://cs.gmu.edu/~sean/book/metaheuristics

Luke S, Cioffi-Revilla C, Panait L, Sullivan K, Balan G (2005) MASON: A multiagent simulation environment. Simulation 81(7):517–527 (DOI:10.1177/0037549705058073)

Magnus PD (2007) Distributed cognition and the task of science. Soc Stud Sci 37(2):297–310 (DOI: 10.1177/0306312706072177, stable JSTOR URL: http://www.jstor.org/stable/25474518)

McClelland JL, Rumelhart DE (1987) Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models. Computational Models of Cognition and Perception. A Bradford book. MIT Press, Cambridge, MA

Muldoon R, Weisberg M (2010) Robustness and idealization in models of cognitive labor. Synthese (online first) (DOI:10.1007/s11229-010-9757-8)

Naveh I, Sun R (2006) A cognitively based simulation of academic science. Comput Mathematical Organization Theory 12(4):313–337 (DOI:10.1007/s10588-006-8872-z)

Naveh I, Sun R (2007) Cognitively based modeling of scientific productivity. In: Rennard J (ed) Handbook of research on nature-inspired computing for economics and management. Idea Group Reference, Hershey, PA, Chap 11, pp 141–152 (DOI:10.4018/978-1-59140-984-7.ch011), also available online at the URL: http://www.cogsci.rpi.edu/~rsun/folder-files/naveh-sun-cmot2006.pdf

Nikolai C, Madey G (2009) Tools of the trade: A survey of various agent based modeling platforms. J Artif Societ Soc Simul 12(2):2, available online at the URL: http://jasss.soc.surrey.ac.uk/12/2/2.html

Pesavento U (1995) An implementation of von Neumann's self-reproducing machine. Artif Life 2(4):337–354 (DOI:10.1162/artl.1995.2.337).

Popper KR (1959) The logic of scientific discovery. Hutchison, London. Translation of: Popper K (1935) Logik der Forschung: Zur Erkenntnistheorie der modernen Naturwissenschaft. Schriften zur Wissenschaftlichen Weltauffassung, vol 9. Julius Springer, Wien

Powell WB (2007) Approximate dynamic programming: Solving the curses of dimensionality. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ

Price DJ de Solla (1963) Little science, big science. Columbia University Press, New York, NY

Pyka A, Gilbert N, Ahrweiler P (2007) Simulating knowledge-generation and distribution processes in innovation collaborations and networks. Cybern Syst Int J 38(7):667–693 (DOI:10.1080/01969720701534059)

Pyka A, Gilbert N, Ahrweiler P (2009) Agent-based modelling of innovation networks – the fairytale of spillover. In: Pyka A, Scharnhorst A (eds) Innovation Networks. New Approaches in Modelling and Analyzing. Springer Complexity: Understanding Complex Systems. Springer, Dordrecht, pp 101–126 (DOI:10.1007/978-3-540-92267-4_5)

Pyka A, Küppers G, Ahrweiler P (2002) Simulating innovation networks. In: Pyka A, Küppers G (eds) Innovation networks: Theory and practice. New Horizons in the Economics of Innovation. Edward Elgar Publishing, Cheltenham

Reynolds C (1987) Flocks, herds and schools: A distributed behavioral model. ACM SIGGRAPH Comput Graph 21(4):25–34 (DOI:10.1145/37401.37406). The article is part of: Stone MC (ed) SIGGRAPH '87: Proceedings of the 14th annual conference on computer graphics and interactive techniques. ACM SIGGRAPH Comput Graph 21(4):1–320

Sawaragi Y, Nakayama H, Tanino T (1985) Theory of multiobjective optimization. Mathematics in Science and Engineering, vol 176. Academic Press, Orlando, FL

Schelling TC (1969) Models of segregation. Am Econ Rev 59(2):488–493 (stable JSTOR URL: http://www.jstor.org/stable/1823701). The article is part of: Papers and proceedings of the one eighty-first annual meeting of the American Economic Association, Chicago, Illinois, December 28–30, 1968. Am Econ Rev 56(2):1–569

Schelling TC (1971a) Dynamic models of segregation. J Math Sociol 1(2):143–186 (DOI:10.1080/0022250X.1971.9989794)

Schelling TC (1971b) On the ecology of micromotives. Public Interest 25(Fall 1971):59–98, available online at the URL: http://pao.chadwyck.co.uk/articles/displayItem.do?QueryType=articles&QueryIndex=journal&ResultsID=12FF8873E3599A75E&ItemNumber=5&BackTo=journalid&BackToParam=QueryType=journals|ItemID=g035|issue=25\(1971:Fall)&journalID=g035

Simon HA (1955) On a class of skew distribution functions. Biometrika 42(3–4):425–440 (DOI:10.1093/biomet/42.3-4.425, stable JSTOR URL: http://www.jstor.org/stable/2333389)

Steuer E (1986) Multiple Criteria Optimization: Theory, Computations, and Application. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, NY

Strevens M (2003) The role of the priority rule in science. J Philos 100(2):55–79 (stable JSTOR URL: http://www.jstor.org/stable/3655792)

Sun R (2002) Duality of the mind: A bottom-up approach toward cognition. Lawrence Erlbaum Associates, Mahwah, NJ

Sun R (2003) A tutorial on CLARION 5.0. Technical report. Rensselaer Polytechnic Institute, Cognitive Science Department, Troy, NY, available online at the URL http://www.cogsci.rpi.edu/$\sim$rsun/sun.tutorial.pdf

Sun R (2006) The CLARION cognitive architecture: Extending cognitive modeling to social simulation. In: Sun R (ed) Cognition and multi-agent interaction: From cognitive modeling to social simulation. Cambridge University Press, Cambridge, Chap 4, pp 79–100 (DOI:10.1017/CBO9780511610721.005)

Sun R, Naveh I (2009) Cognitive simulation of academic science. In: The 2009 international joint conference on neural networks: IJCNN 2009 conference proceedings, Atlanta, Georgia, USA, June 14–19, 2009. IEEE Computer Society Press, Piscataway, NJ, pp 3011–3017 (DOI:10.1109/IJCNN.2009.5178638)

Thagard P (1993a) Computational philosophy of science. A Bradford book. MIT Press, Cambridge, MA

Thagard P (1993b) Societies of minds: Science as distributed computing. Stud Hist Philos Sci Part A 24(1):49–67 (DOI:10.1016/0039-3681(93)90024-E)

Thagard P, Holyoak K (1985) Discovering the wave theory of sound: Inductive inference in the context of problem solving. In: Joshi A (ed) IJCAI 85: Proceedings of the ninth international joint conference on artificial intelligence, August 18–23, 1985, Los Angeles, California. Morgan-Kaufmann Publishers, San Francisco, CA, vol 1, pp 610–612, available online at the URL: http://ijcai.org/Past\Proceedings/IJCAI-85-VOL1/PDF/117.pdf

Toulmin S (1972) Human understanding: The collective use and evolution of concepts. Princeton University Press, Princeton, NJ von Neumann J (1966) Theory of self-reproducing automata. University of Illinois Press, Urbana, IL

Weisberg M, Muldoon R (2009) Epistemic landscapes and the division of cognitive labor. Philos Sci 76(2):225–252 (DOI:10.1086/644786, stable JSTOR URL: http://www.jstor.org/stable/10.1086/644786)

Zollman KJS (2007) The communication structure of epistemic communities. Philos Sci 74(5):574–587 (DOI:10.1086/525605, stable JSTOR URL: http://www.jstor.org/stable/10.1086/525605http://www.jstor.org/stable/10.1086/525605)

AUTHOR QUERIES

AQ1. Please check whether the inserted citation for Table 4.1 is appropriate.

AQ2. Please check whether the inserted citation for Figs. 4.1–4.10, 4.12–4.14, 4.16 is appropriate.

AQ3. Please check whether the inserted citation for Tables 4.2 and 4.3 are appropriate.

Chapter 5
Evolutionary Game Theory and Complex Networks of Scientific Information

Matthias Hanauske

5.1 Introduction

The encounter of information science with the theory of complex networks is the main characteristic of a realistic model of science dynamics. Complex information networks and the social dimension of the network of researchers are combined in a multi-level network model which functions as the topological background of the whole market of scientific information. A main goal of academic research is the diffusion of new research results. This is achieved by interaction between scientists through reading and citing other authors' work (Bernius et al. 2010). Complex citation, co-authorship, and semantic networks have been evolved in reality, and the theoretical description of the dynamical behavior of these networks has been addressed in several chapters of this book. The evolution of the market of scientific information depends not only upon the researchers' actions, but also upon the actions of other actors involved in the knowledge-creation process (journals, libraries, funding agencies, etc.). For some years, the market of scientific publishing has been forced to make major changes in the process of distributing research results among scientists. First, the increase in digitalization brought a shift towards electronic publication, and second, shrinking library budgets in combination with a constant rise of journal prices have resulted in massive cancellations of journal subscriptions. In order to regain broad access to research findings, alternative ways of publishing scientific literature have been developed and have received increased attention. These new models are summarized under the term "Open Access (OA)" AQ1 (Bernius and Hanauske 2007) (Table 5.1).

Within this chapter, the market of scientific information is modeled as a game between various actors involved in the knowledge-creation process. The main

---

M. Hanauske (✉)
Institute of Information Systems, Goethe-University, Grüneburgplatz 1, 60323 Frankfurt/Main
e-mail: hanauske@wiwi.uni-frankfurt.de

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_5,

Table 5.1 Major questions raised in this chapter and their answers

Major questions raised in this chapter	And their answers
1. Why should I deal with game theory?	By analysing the game structure of a specific decision problem, decision-makers can learn a lot about the problem they are involved in.
2. What is the difference between game theory and evolutionary game theory?	Evolutionary game theory uses game-theoretical concepts, but focuses on the strategic decisions within a whole population of players, and describes the evolutionary, time-dependent dynamics of the population.
3. What do I need for a game-theoretical analysis of my specific decision problem?	You need only three things: The set of players, the set of available actions (strategies), and the payoff structure of the underlying game.
4. What are Nash equilibria, dominant strategies, and evolutionary stable strategies?	These different equilibrium concepts will be defined, visualized, and explained in detail (see Sect. 5.2). They are, for example, important for the definition of different game classes.
5. What types of games are possible?	Symmetric and unsymmetric games. For symmetric games, the three different game classes – "dominant games," "coordination games," and "anti-coordination games," are possible. For unsymmetric games, there are three major categories possible: "corner class," "saddle class," and "center class".
6. How can evolutionary game theory be applied to science dynamics?	Two applications are discussed within this chapter. Section 5.3.1: "Scientific communication and the open access decision" and Sect. 5.3.2: "Evolution of Hub-and-Spoke Communication Networks".
7. In the future, will scientific information be free of charge for everyone?	Scientists face a dilemma: Considering a potential loss in reputation, incentives to perform open access are missing (see Sect. 5.3.1). Scientific publishers also face a dilemma, as they fear a profit loss within a totally "green-open-access publishing market" (see Sect. 5.3.2).
8. Evolutionary game theory depends only on a few open parameters. How can that be? Isn't nature very complicated?	With the use of this simple model, one can learn a lot about the underlying game. However, some aspects are not included within classical evolutionary game theory. Some amplifications of the classical theory ("Evolutionary Game Theory on Complex Networks" and "Evolutionary Quantum Game Theory") are discussed in Sect. 5.4.

research goal of the chapter is to understand different publication norms within the scientific community, especially the description of the time evolution of the average strategic decision of different actor populations, using the framework of the evolutionary game theory. How can one include group behavior and social norms (which might be caused by cultural or moral standards) into the theory of population dynamics formulated within the evolutionary game theory? Evolutionary game theory on complex networks using agent-based computation methods and quantum game theory are recently developed models, and they will be discussed briefly at the end of this chapter (see Sect. 5.4).

Within this chapter of the book *Models of Science Dynamics–Encounters between Complexity Theory and Information Science*, the framework of evolutionary game theory (EGT) is described in detail. After a general introduction and a discussion of a simple game-theoretical example, the grounding of EGT (Sect. 5.2) and a brief literature review is presented. The formal mathematical model, different concepts of equilibria, and various classes of evolutionary games will be defined, explained, and visualized. In Sect. 5.3, two applications are presented. The first one (see Sect. 5.3.1) focuses on the open-access game of scientific communication and extends it to an evolutionary game (for details, see (Hanauske et al. 2007, 2010b)). The second application (see Sect. 5.3.2) focuses on the evolution of the interconnected network of scientific journals and scientific authors within a formal "Hub-and-Spoke Communication Network" model. The combination of evolutionary game theory with the theory of complex networks and the description of a new framework that includes group behavior and social norms into evolutionary population dynamics are briefly explained in Sect. 5.4. The chapter ends with a short summary.

5.2 Evolutionary Game Theory

In 1928, the main inventor of game theory – Johann (John) von Neumann – published the first article on this important topic (von Neumann 1928).[1] The first book about game theory was published in 1944 by von Neumann and Morgenstern (von Neumann and Morgenstern 1944). Evolutionary game theory (Smith and Price 1973; Smith 1974, 1982; Schlee 2004; Miekisz 2008; Szabó and Fáth 2007; Schlee 2004; Amann 1999; Hanauske 2009) was developed after J.M. Smith had found that stationary solutions to evolutionary differential equations are connected with game theory (Smith 1972). In the following years, applications in respect to biological systems (Sinervo and Lively 1996; Turner and Chao 1999; Kerr et al. 2002; Fraser et al. 2002; Nowak and Sigmund 2002, 2003) and socio-economic systems–e.g., "public good" games (Clemens and Riechmann 2006), cultural or moral developments (Enquist and Ghirlanda 2007; Harms and Skyrms 2008), the evolution of languages (Pawlowitsch 2007), social learning (Enquist and Ghirlanda 2007), the evolution of social norms (Axelrod 1997; Ostrom 2000), the financial crisis (Hanauske et al. 2009), and the evolution of social networks (Szabó and Fáth 2007; Janssen and Ostrom 2006; Ostrom 2009) – came into the focus of research.

---

[1] In principle, the groundings of GT go back to 1800 (e.g. Antoine-Augustin Cournot (1801–1877) and Francis Ysidro Edgeworth (1845–1926) (Söllner 2001)). Additionally, in the 1913, Ernst Zermelo had discussed the chess game using a backward-induction method (Zermelo 1913). However, the first formal, mathematical description of GT was developed by John von Neumann in the year 1928 (von Neumann 1928).

5.2.1 Game Theory: A Simple Example

The necessary definitions and fundamental basics of GT and EGT will be explained in the next subsection; however, the following section explains the use of game-theoretical concepts with one simple example:

Two persons (Emma and Hans) have to make a decision. Each of them has to choose between two possible actions. For both of them it is an important decision, as they might get a great benefit (or a punishment) if they choose the "right" (or "wrong") decision. The amount of the potential benefit depends on the decisions of both persons and not only on the action of one. Unfortunately, they do not have any possibility of communicating with the other one to coordinate their actions.

GT is a mathematical concept used to analyze such decision states. Every quantitative mathematical model that tries to explain processes happening in nature begins with a definition of the necessary parameters. In the following, the parameter A or B (later also μ) will be used to describe a person, a player, a decision-maker, or even a firm or an animal. In the above example, the parameter A means "Emma" and the parameter B means "Hans". The parameter $\mathcal{S}^A$ will be used to describe the set of possible strategies (actions) available to Emma, whereas $\mathcal{S}^B$ describes the set of available actions of player "Hans." In the above example, this would be written as $\mathcal{S}^A = \left\{s_1^A, s_2^A\right\}$, as Emma can only choose between two possible actions namely, strategy one (s_1^A) and strategy two (s_2^A). The strategy space of Hans is written in a similar form: $\mathcal{S}^B = \left\{s_1^B, s_2^B\right\}$. The parameter U is used to quantify the potential benefit (or the amount of punishment) given to players after they have announced their final decisions.

In principle, to define a game Γ, one needs three things:

- Who is playing the game? Definition of the set of players: $\mathcal{I} = \{A, B, \ldots,\} = \{\text{Emma}, \text{Hans}, \ldots,\}$
- What can the players do? Definition of the set of actions (strategies) available for each player: $\mathcal{S}^A = \mathcal{S}^{\text{Emma}} = \left\{s_1^A, s_2^A, \ldots,\right\}$ and $\mathcal{S}^B = \mathcal{S}^{\text{Hans}} = \left\{s_1^B, s_2^B, \ldots,\right\}$
- How much can the players win or lose? Definition of the payoff structure of the game: $\hat{\mathcal{U}}^A = \hat{\mathcal{U}}^{\text{Emma}}$ and $\hat{\mathcal{U}}^B = \hat{\mathcal{U}}^{\text{Hans}}$

Every decision-maker who wants to analyse her/his decision problem (her/his game) with game-theoretical concepts has to define these three things – therefore, the simple example is extended with the use of an additional little story. The binary decision of Emma (Player A) and Hans (Player B) could be "To stay" or "To go," or it could be simply to choose between two strategies (e.g., {*buy, don't buy*}, {*left, right*}, $\{above, below\}$, $\{s_1, s_2\}$). The benefit if both choose the strategy s_1 is very good for both of them, and the parameter U_{11} is used in the following to quantify this benefit. If Emma and Hans choose the strategy s_2, it will be bad for both of them, and the parameter U_{22} quantifies the value of punishment for both players. If Emma decides to stay (s_1^A) and Hans goes, the outcome for Hans will be even slightly better than the situation for him if both stay ($U_{11}^B < U_{12}^B$); the same holds true for Emma: ($U_{11}^A < U_{21}^A$). However, if Emma chooses the strategy s_2^A and Hans stays (strategy

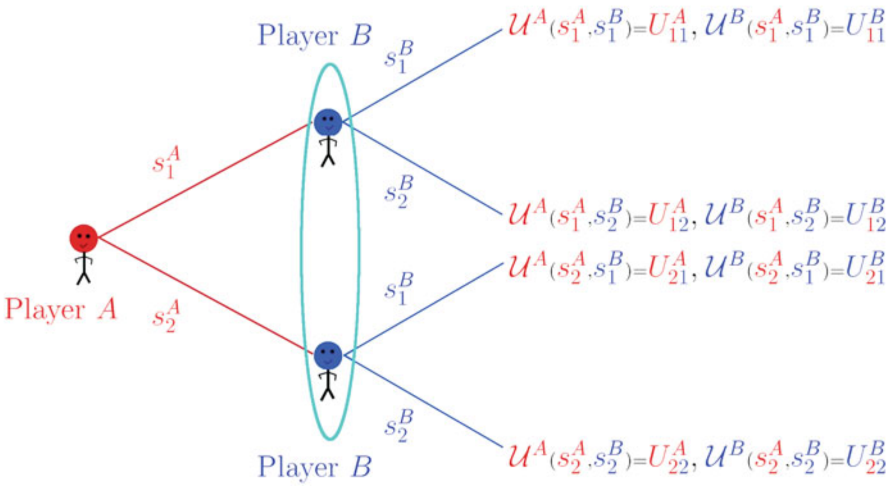


Fig. 5.1 Game tree of a (2 person)–(2 strategy) game with payoff for player A ($\mathcal{U}^A$) and player B ($\mathcal{U}^B$)

s_1^A), the outcome for Hans will be extremely bad ($U_{21}^B \ll U_{22}^B$); the same holds true for Emma: ($U_{12}^A \ll U_{22}^A$). Figure 5.1 visualizes this (two player)–(two strategy) game as a game tree with four possible payoff outcomes.

GT analyses such decision states, using mathematically defined equilibrium concepts. The most famous concept of this kind is called the "Nash equilibrium" (NE). As player B does not know for sure what player A will do, he starts to think what would be the best for him, if player A chose the strategy s_1^A (staying): *"It would be good for me if player A stays and I stay, but in this case it would be even better for me to go."* After remaining a moment at this thought, player B starts to think in the other direction: *"If player A goes and I stay, it will be extremely bad for me – it is really advisable for me to go!"* Within the framework of classical GT, the predicted outcome of this example is that both players decide to go. In the language of game theory, the strategy s_2 is the only NE of this example, and as the game is a (two player)–(two strategy) normal-form game, s_2 is even a dominant strategy. To be more precise:

The strategy combination (s_2^A, s_2^B) is a Nash equilibrium because:

$$\text{Nash equilibrium at } (s_2^A, s_2^B):$$
$$\begin{aligned} \mathcal{U}^A(s_2^A, s_2^B) &= U_{22}^A \;\geq\; \mathcal{U}^A(s^A, s_2^B) \quad \forall\; s^A \in \mathcal{S}^A = \left\{s_1^A, s_2^A\right\} \\ \mathcal{U}^B(s_2^A, s_2^B) &= U_{12}^B \;\geq\; \mathcal{U}^B(s_2^A, s^B) \quad \forall\; s^B \in \mathcal{S}^B = \left\{s_1^B, s_2^B\right\} \end{aligned} \tag{5.1}$$

The tragedy of this game is that after both players have made their decision, they are in a worse situation than when they had chosen the strategy s_1 ($U_{22}^A < U_{11}^A$ and $U_{22}^B < U_{11}^B$) – therefore, the game belongs formally to the class of prisoner's dilemma games (class of dominant games with a dilemma).

Depending on the payoff structure of the game ($\hat{\mathcal{U}}^A$ and $\hat{\mathcal{U}}^B$), different game classes and outcomes are possible. By analysing the game structure of a specific decision problem, decision makers can learn a lot about the problem they are involved in.

The simple example within this subsection was used to explain game-theoretical concepts. EGT uses these concepts, but focuses on the strategic decisions within a whole population of players. There exist not only one Emma and one Hans, but a whole group of players like Emma (group A) and a whole group of players like Hans (group B). They do not play the game only once – at each time increment the Emma's and the Hans's come together, play the game, receive their payoffs, and search the next game partner for the next time increment. The framework of EGT only needs one piece of additional information about the game Γ: What is the fraction of players within group A (group A) choosing strategy s_1^A (choosing strategy s_1^B) at time zero – the initial value of the strategic decision of the whole population. Knowing the game Γ and the initial value, the framework of EGT is able to show the evolutionary dynamics of the population, and it gives answers about the thing everybody wants to know: "How is it going to the end?"

5.2.2 *Definition and Key Aspects of Evolutionary Game Theory*

EGT is a time-dependent dynamical extension of "Game Theory" (GT), which itself is a mathematical toolbox to explain interdependent decision processes happening in biological or socio-economic systems. As the variety of different concepts in GT is very large, and the article is not meant to summarize only GT, the game-theoretical concepts presented in this article will only focus on "strategic-form games",[2] and the article does not discuss "extensive-form games" nor "cooperative games." In the following, the formal framework of the mixed extension of a (N player)–(m strategy) game in strategic form will be defined:

N-person game: $\Gamma := \left(\mathcal{I}, \tilde{\mathcal{S}}, \tilde{\mathcal{U}}\right)$

Set of players: $\mathcal{I} = \{1, 2, \ldots, N\}$

Pure strategy space: $\mathcal{S} = \mathcal{S}^1 \times \mathcal{S}^2 \times \ldots \times \mathcal{S}^N$

Pure strategy space of player $\mu \in \mathcal{I}$: $\mathcal{S}^\mu = \left\{(s_1^\mu, s_2^\mu, \ldots, s_{m_\mu}^\mu)\right\}$

Mixed-strategy space: $\tilde{\mathcal{S}} = \tilde{\mathcal{S}}^1 \times \tilde{\mathcal{S}}^2 \times \ldots \times \tilde{\mathcal{S}}^N$

Mixed-strategy space of player $\mu \in \mathcal{I}$:

$$\tilde{\mathcal{S}}^\mu = \left\{(\tilde{s}_1^\mu, \tilde{s}_2^\mu, \ldots, \tilde{s}_{m_\mu}^\mu) \mid \sum_{i=1}^{m_\mu} \tilde{s}_i^\mu = 1, \tilde{s}_i^\mu \geq 0, i = 1, 2, \ldots, m_\mu\right\} \tag{5.2}$$

---

[2] The category of "strategic-form games" is often also called "non-cooperative games".

Number of strategies available for player $\mu \in \mathcal{I} \quad m_\mu$

Mixed-strategy profile of player $\mu \in \mathcal{I}$: $\quad \tilde{s}^\mu = \left(\tilde{s}^\mu_1, \tilde{s}^\mu_2, \dots, \tilde{s}^\mu_{m_\mu}\right)^T \in \tilde{\mathcal{S}}^\mu$

Vector function of mixed payoffs: $\quad \tilde{\mathcal{U}} = \left(\tilde{\mathcal{U}}^1, \tilde{\mathcal{U}}^2, \dots, \tilde{\mathcal{U}}^N\right) : \tilde{\mathcal{S}} \to \mathbb{R}^N$

Mixed payoff for player $\mu \in \mathcal{I}$:

$$\tilde{\mathcal{U}}^\mu(\tilde{s}^1, \tilde{s}^2, \dots, \tilde{s}^N) = \sum_{i_1=1}^{m_1} \sum_{i_2=1}^{m_2} \dots \sum_{i_N=1}^{m_N} \mathcal{U}^\mu(s^1_{i_1}, s^2_{i_2}, \dots, s^N_{i_N}) \prod_{\nu=1}^{N} \tilde{s}^\nu_{i_\nu}$$

Definition (5.2) expresses that three main quantities are necessary to define a (N player)–(m strategy) game in strategic form. The first quantity, the set of players $\mathcal{I}$, includes all of the actors involved in the underlying game. In respect to the focus of this book, $\mathcal{I}$ could be understood as the set of entities involved in the knowledge-creation process (subsets of $\mathcal{I}$: researchers, journals, libraries, funding agencies, etc.). The second quantity, the set of pure strategies $\tilde{\mathcal{S}}$, expresses all of the available strategies of all of the actors involved in the game. In principle, each actor $\mu \in \mathcal{I}$ could have her/his own set of available strategies ($\mathcal{S}^\mu$). If we focus again on a model of science, the different subgroups of $\mathcal{I}$ will have similar strategy spaces (strategy space of scholars, strategy space of journals, etc.). The set of mixed strategies of player μ ($\tilde{\mathcal{S}}^\mu$) is a mathematical amplification of the set of pure strategies ($\tilde{\mathcal{S}}^\mu$). The elements belonging to the set of mixed strategies ($\tilde{\mathbf{s}}^\mu = (\tilde{s}^\mu_1, \tilde{s}^\mu_2, \dots, \tilde{s}^\mu_{m_\mu}) \in \mathcal{S}^\mu$) consist of m_μ real numbers ($\tilde{s}^\mu_i \in [0,1]\ \forall\, i \in \{1, 2, \dots, m_\mu\}$) and can be interpreted as the probability of player μ for choosing the pure strategy s^μ_i. The third quantity, the mixed strategy payoff function $\tilde{\mathcal{U}}$, is used to quantify the potential benefit (or the amount of punishment) given to the persons. The amount of the potential benefit (punishment) given to a player μ ($\tilde{\mathcal{U}}^\mu$) depends on the actions of all players and not only on the strategy decision of player μ.

To be more precise, the following part is constrained to the strategic form of an unsymmetric (or symmetric) (2 player)–(2 strategy) game Γ (for details, see (Hanauske 2009; Szabó and Fáth 2007)):

(2 × 2) game: $\quad \Gamma := \left(\{A, B\}, \mathcal{S}^A \times \mathcal{S}^B, \hat{\mathcal{U}}^A, \hat{\mathcal{U}}^B\right)$

Set of pure strategies of player A and B:

$\mathcal{S}^A = \left\{s^A_1, s^A_2\right\}, \ \mathcal{S}^B = \left\{s^B_1, s^B_2\right\}$

Set of mixed strategies of player A and B:

$\tilde{\mathcal{S}}^A = \left\{\tilde{s}^A_1, \tilde{s}^A_2\right\}, \ \tilde{\mathcal{S}}^B = \left\{\tilde{s}^B_1, \tilde{s}^B_2\right\}$

Mixed payoff of player $\mu \in \{A, B\}$: $\quad \tilde{\mathcal{U}}^\mu : \left(\tilde{\mathcal{S}}^A \times \tilde{\mathcal{S}}^B\right) \to \mathbb{R}$

$\tilde{\mathcal{U}}^\mu((\tilde{s}^A_1, \tilde{s}^A_2), (\tilde{s}^B_1, \tilde{s}^B_2)) = U^\mu_{11}\tilde{s}^A_1\tilde{s}^B_1 + U^\mu_{12}\tilde{s}^A_1\tilde{s}^B_2 + U^\mu_{21}\tilde{s}^A_2\tilde{s}^B_1 + U^\mu_{22}\tilde{s}^A_2\tilde{s}^B_2$

$$\text{Payoff matrix for player A and B:} \quad \hat{\mathcal{U}}^A = \begin{pmatrix} U^A_{11} & U^A_{12} \\ U^A_{21} & U^A_{22} \end{pmatrix}, \ \hat{\mathcal{U}}^B = \begin{pmatrix} U^B_{11} & U^B_{12} \\ U^B_{21} & U^B_{22} \end{pmatrix} \tag{5.3}$$

The set of mixed strategies of player A ($\tilde{\mathcal{S}}^A$) and player B ($\tilde{\mathcal{S}}^B$) is a mathematical amplification of the set of pure strategies ($\mathcal{S}^A$ and $\mathcal{S}^B$). The elements belonging to the set of mixed strategies ($\tilde{\mathbf{s}}^\mu = (\tilde{s}_1^\mu, \tilde{s}_2^\mu) \in \mathcal{S}^\mu$) of player $\mu = A, B$ consist of two real numbers ($\tilde{s}_1^\mu \in [0,1]$ and $\tilde{s}_2^\mu \in [0,1]$) and can be interpreted as the probability of player μ for choosing the strategy 1 ($\tilde{s}_1^\mu$) or 2 ($\tilde{s}_2^\mu$). For two-strategy games, the following normalization condition has to be fulfilled: $\tilde{s}_1^\mu + \tilde{s}_2^\mu = 1 \ \forall \, \mu = A, B$.

Due to the normalizing condition, it is possible to simplify the functional dependence of the mixed-strategy payoff function:

$$\begin{aligned} \tilde{\mathcal{U}}^\mu &: ([0,1] \times [0,1]) \to \mathbb{R} \\ \tilde{\mathcal{U}}^\mu(\tilde{s}^A, \tilde{s}^B) &= U_{11}^\mu \tilde{s}^A \tilde{s}^B + U_{12}^\mu \tilde{s}^A (1 - \tilde{s}^B) + \\ &\quad + U_{21}^\mu (1 - \tilde{s}^A) \tilde{s}^B + U_{22}^\mu (1 - \tilde{s}^A)(1 - \tilde{s}^B) \, , \end{aligned} \tag{5.4}$$

where $\tilde{s}^A := \tilde{s}_1^A$, $\tilde{s}^B := \tilde{s}_1^B$, $\tilde{s}_2^A = 1 - \tilde{s}_1^A$ and $\tilde{s}_2^B = 1 - \tilde{s}_1^B$.

In the following, two fundamental equilibrium concepts are defined, namely the *equilibrium in dominant strategies* and the *Nash equilibrium*.

A strategy combination $(\tilde{s}^{A\dagger}, \tilde{s}^{B\dagger})$ is an equilibrium in dominant strategies if the following conditions are fulfilled:

$$\begin{gathered} \text{Equilibrium in dominant strategies:} \\ \tilde{\mathcal{U}}^\mu(\tilde{s}^{A\dagger}, \tilde{s}^{B\dagger}) \ \geq \ \tilde{\mathcal{U}}^\mu(\tilde{s}^A, \tilde{s}^B) \quad \forall \ \mu = A, B \ \text{and} \quad \tilde{s}^A, \tilde{s}^B \in [0,1] \end{gathered} \tag{5.5}$$

A strategy combination $(\tilde{s}^{A*}, \tilde{s}^{B*})$ is called a Nash equilibrium if:

$$\begin{aligned} \text{Nash equilibrium: } \tilde{\mathcal{U}}^A(\tilde{s}^{A*}, \tilde{s}^{B*}) \ &\geq \ \tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^{B*}) \quad \forall \ \tilde{s}^A \in [0,1] \\ \tilde{\mathcal{U}}^B(\tilde{s}^{A*}, \tilde{s}^{B*}) \ &\geq \ \tilde{\mathcal{U}}^B(\tilde{s}^{A*}, \tilde{s}^B) \quad \forall \ \tilde{s}^B \in [0,1] \end{aligned} \tag{5.6}$$

An interior (mixed-strategy) NE $(\tilde{s}^{A\star}, \tilde{s}^{B\star})$ is a special case of the Definition 5.6, as the partial derivative of the mixed-strategy payoff function vanishes at the value of the interior NE:

$$\begin{gathered} \text{Interior Nash equilibrium:} \\ \left. \frac{\partial \tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)}{\partial \tilde{s}^A} \right|_{\tilde{s}^B = \tilde{s}^{B\star}} = 0 \quad \forall \ \tilde{s}^A \in [0,1] \ , \ \tilde{s}^{B\star} \in]0,1[\\ \left. \frac{\partial \tilde{\mathcal{U}}^B(\tilde{s}^A, \tilde{s}^B)}{\partial \tilde{s}^B} \right|_{\tilde{s}^A = \tilde{s}^{A\star}} = 0 \quad \forall \ \tilde{s}^B \in [0,1] \ , \ \tilde{s}^{A\star} \in]0,1[\end{gathered} \tag{5.7}$$

The defined equilibrium concepts will be used in Sect. 5.2.3 to classify games into different classes. The hitherto defined mathematical constructs can be used to analyze one-shot (2×2) games, while the following equations will describe the time evolution of the strategic behavior of a large group of players (population). At each time increment all of the individual players of the population search randomly for a partner to play a (2×2) game. Then, after the players have chosen their strategies and have received their payoffs, they search again for the next game partner. To describe the time evolution of such a repeated version of the game Γ, *replicator dynamics* has been developed. As the payoff matrices ($\hat{\mathcal{U}}^A$ and $\hat{\mathcal{U}}^B$) of the two persons playing the game are in general unsymmetric, the whole population of players separates into the two subpopulations "A" and "B." Replicator dynamics, formulated within a system of differential equations, defines in which way the population vector $\mathbf{x}^\mu = (x_1^\mu, x_2^\mu)$ evolves in time. Each component $x_i^\mu = x_i^\mu(t)$ ($i = 1, 2$ and $\mu = A, B$) describes the time evolution of the fraction of different player types i in the μ-subpopulation, where a type-i player is understood as an actor μ playing strategy s_i^μ. Similar to the normalizing condition of the mixed strategies, the two population vectors $\mathbf{x}^A$ and $\mathbf{x}^B$ have to fulfill the normalizing conditions of a unity vector:

$$x_i^\mu(t) \geq 0 \quad \text{and} \quad \sum_{i=1}^{2} x_i^\mu(t) = 1 \quad \forall\, i = 1, 2\,,\ t \in \mathbb{R},\ \mu = A, B. \tag{5.8}$$

The structure of the time evolution of the components of the two population vectors $\mathbf{x}^A(t) = (x_1^A(t), x_2^A(t)$ and $\mathbf{x}^B(t) = (x_1^B(t), x_2^B(t))$ is formulated through a system of differential equations, known as the equation of *Replicator Dynamics* (Amann 1999; Schlee 2004; Miekisz 2008; Hanauske 2009; Szabó and Fáth 2007):

$$\begin{aligned}
\frac{dx_i^A(t)}{dt} &= x_i^A(t) \left[\sum_{l=1}^{2} U_{il}^A\, x_l^B(t) - \sum_{l=1}^{2} \sum_{k=1}^{2} U_{kl}^A\, x_k^A(t)\, x_l^B(t) \right] \\
\frac{dx_i^B(t)}{dt} &= x_i^B(t) \left[\sum_{l=1}^{2} U_{li}^B\, x_l^A(t) - \sum_{l=1}^{2} \sum_{k=1}^{2} U_{lk}^B\, x_l^A(t)\, x_k^B(t) \right]
\end{aligned} \tag{5.9}$$

As the number of available strategies in our approach is restricted to two, it is possible to substitute the second strategy by using condition 5.8: $x_2^A = 1 - x_1^A$ and $x_2^B = 1 - x_1^B$. The system of differential equations (5.9) can therefore be formulated as follows ($x(t) := x_1^A(t)$, $y(t) := x_1^B(t)$):

$$\begin{aligned}
\frac{dx(t)}{dt} &= \Bigg(\underbrace{U_{11}^A - U_{21}^A}_{:=a^A} + \underbrace{U_{22}^A - U_{12}^A}_{:=b^A} \Bigg) \left(x(t) - (x(t))^2 \right)\, y(t) \\
&\quad - \underbrace{\left(U_{22}^A - U_{12}^A \right)}_{:=b^A} \left(x(t) - (x(t))^2 \right) \\
&= \left(a^A + b^A \right) \left(x(t) - (x(t))^2 \right)\, y(t) - b^A \left(x(t) - (x(t))^2 \right) =: g_A(x, y)
\end{aligned}$$

$$\begin{aligned}\frac{dy(t)}{dt} &= \Bigg(\underbrace{U^B_{11} - U^B_{21}}_{:=a^B} + \underbrace{U^B_{22} - U^B_{12}}_{:=b^B}\Bigg)\left(y(t)-(y(t))^2\right)x(t)\\ &\quad - \underbrace{\left(U^B_{22}-U^B_{12}\right)}_{:=b^B}\left(y(t)-(y(t))^2\right)\\ &= \left(a^B+b^B\right)\left(y(t)-(y(t))^2\right)x(t) - b^B\left(y(t)-(y(t))^2\right) =: g_B(x,y)\end{aligned} \qquad (5.10)$$

Equation (5.10) describes the time evolution of the strategic behavior of two separate subpopulations playing an unsymmetric bimatrix game. The fraction of players choosing strategy s_1 at time t of the subpopulation "A" is quantified by $x(t)$, whereas $y(t)$ describes the average strategic choice of subpopulation "B." The time evolution of the coupled system of differential equations (5.10) depends on the properties of the two functions $g_A(x,y)$ and $g_B(x,y)$ and on the initial conditions $x(t=0)$ and $y(t=0)$.

If we focus on a model of science, the two different subpopulations playing the evolutionary game could be, for example, the group of scholars (subpopulation "A") and the group of journals (subpopulation "B"). The two pure strategies of a member of the group A of researchers could be based on any relevant, recurring binary decision a scholar has to decide during her/his research lifetime (e.g., does she/he want to put her/his new article on a open-access repository). The two pure strategies of a member of the group B of journals could be any recurring binary decision a journal has to make (e.g., does the journal allow the authors to put their submitted article version on an open-access repository). The fraction of researchers choosing strategy $s_1^A \hat{=}$ (put the article on an open-access repository) at time t is quantified by $x(t)$, where $x = 1$ corresponds to a situation where every scholar uses open-access repositories, and $x = 0$ means nobody uses them. Similarly, the fraction of journals choosing strategy $s_1^A \hat{=}$ (allowing open-access repositories) at time t is quantified by $y(t)$, where $y = 1$ corresponds to a situation where every journal allows open-access repositories and $y = 0$ means no journal allows it. The two payoff matrices finally quantify the potential benefit to the researchers ($\hat{\mathcal{U}}^A$) and journals ($\hat{\mathcal{U}}^B$). This particular bimatrix game will be discussed in more detail within Sect. 5.3.2.

By restricting the underlying payoff matrix to be symmetric ($\hat{\mathcal{U}}^A \equiv \left(\hat{\mathcal{U}}^B\right)^T$, $U_{lk} := U^A_{lk} = U^B_{kl}$), the two separate subpopulations (A and B) cannot be distinguished any more and the system of differential equations (5.9) simplifies as follows:

$$\begin{aligned}\frac{dx^A_i(t)}{dt} &= x^A_i(t)\left[\sum_{l=1}^{2} U_{il}\, x^B_l(t) - \sum_{l=1}^{2}\sum_{k=1}^{2} U_{kl}\, x^A_k(t)\, x^B_l(t)\right]\\ \frac{dx^B_i(t)}{dt} &= x^B_i(t)\left[\sum_{l=1}^{2} U_{il}\, x^A_l(t) - \sum_{l=1}^{2}\sum_{k=1}^{2} U_{kl}\, x^A_l(t)\, x^B_k(t)\right]\end{aligned} \qquad (5.11)$$

Equation (5.11) indicates that the mathematical structures of the two population vectors $\mathbf{x}^A$ and $\mathbf{x}^B$ are identical, which simply means that a symmetric evolutionary game can be described by a single population vector $\mathbf{x} := \mathbf{x}^A = \mathbf{x}^B$. In respect to a model of science, this means that (5.11) can only be used for subgames with strategic decisions involving only one set of knowledge entities. Therefore the system of differential equations (5.11) reduces to one single equation:

$$\frac{dx_i(t)}{dt} = x_i(t) \left[\underbrace{\sum_{l=1}^{2} U_{il}\, x_l(t)}_{:= f_i(t)} - \underbrace{\sum_{l=1}^{2} \sum_{k=1}^{2} U_{kl}\, x_k(t)\, x_l(t)}_{:= \bar{f}(t)} \right] \tag{5.12}$$

where $f_i(t)$ is the fitness of type i and $\bar{f}(t) = \sum_{i=1}^{2} f_i(t)$ is the average fitness of the whole population. Again, the overall vector $\mathbf{x} = (x_1(t), x_2(t))$ has to fulfill the normalizing conditions of a unity vector:

$$x_i(t) \geq 0 \;\; \forall\, i = 1,2 \quad \text{and} \quad \sum_{i=1}^{2} x_i(t) = 1 \;\; \forall\, t \in \mathbb{R}. \tag{5.13}$$

For a symmetric game, (5.12) can therefore be simplified as follows:

$$\begin{aligned}
\frac{dx}{dt} &= x \left[U_{11}(x - x^2) + U_{12}(1 - 2x + x^2) + U_{21}(x^2 - x) + U_{22}(2x - x^2 - 1) \right] \\
&= x \left[\underbrace{(U_{11} - U_{21})}_{:=a}(x - x^2) - \underbrace{(U_{22} - U_{12})}_{:=b}(1 - 2x + x^2) \right] \\
&= x \left[a(x - x^2) - b(1 - 2x + x^2) \right] \\
&=: g(x) \text{with:} \quad x = x(t) := x_1(t) \quad \text{and} \quad x_2(t) = (1 - x(t))
\end{aligned} \tag{5.14}$$

The function $x(t)$, describing the fraction of players choosing the strategy s_1 at time t, depends on the function $g(x)$ and on the initial starting value $x(t = 0)$. The stationary solution of the asymptotic behavior $\lim_{t \to \infty}(x(t))$ depends also on $g(x)$ and on the initial condition, and it is formalized within the mathematical concept of the *Evolutionary Stable Strategy* (ESS). For a general 2-player game Γ with the mixed payoff functions $\tilde{\mathcal{U}}^A$ and $\tilde{\mathcal{U}}^B$, a strategy combination $(\tilde{s}^{A*}, \tilde{s}^{B*}) \in ([0,1] \times [0,1])$ is defined as an (ESS) if:

a) $(\tilde{s}^{A*}, \tilde{s}^{B*})$ is a Nash equilibrium of the game

b) $\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B) \leq \tilde{\mathcal{U}}^A(\tilde{s}^{A*}, \tilde{s}^B) \qquad \forall \quad \tilde{s}^A \in r^A(\tilde{s}^{B*}), \;\; \tilde{s}^B \neq \tilde{s}^{B*}$
$\tilde{\mathcal{U}}^B(\tilde{s}^A, \tilde{s}^B) \leq \tilde{\mathcal{U}}^B(\tilde{s}^A, \tilde{s}^{B*}) \qquad \forall \quad \tilde{s}^B \in r^B(\tilde{s}^{A*}), \;\; \tilde{s}^A \neq \tilde{s}^{A*}.$

Let $r^B(\tilde{s}^A)$ and $r^A(\tilde{s}^B)$ signify the best response functions of players B and A to the strategy $\tilde{s}^A$ and $\tilde{s}^B$, respectively. An ESS $(\tilde{s}^{A*}, \tilde{s}^{B*})$ therefore needs to be a Nash equilibrium of the game, and also the inequations b) should be fulfilled for any strategy combination $(\tilde{s}^A, \tilde{s}^B)$ belonging to the set of best responses to $(\tilde{s}^{A*}, \tilde{s}^{B*})$.

This survey has focused on deterministic evolutionary game dynamics and has specially concentrated on replicator dynamics. Stochastic evolutionary game dynamics and adaptive or rational learning processes have not been discussed (for a detailed analysis, see e.g., Sandholm 2010). The discussed evolutionary dynamics uses only the revision protocol of replicator dynamics and other possible types of dynamics (nonlinear payoff functions, general imitation dynamics, best-response dynamics, logit dynamics and Brown-von Neumann–Nash dynamics) were not discussed within this chapter either (for a detailed analysis, see e.g., (Sandholm 2010; Hofbauer and Sigmund 2003)). The conjunction of evolutionary game theory with the theory of complex networks using concepts from agent-based modeling is a new and interesting scientific topic, but it is not addressed within this chapter (for a detailed analysis, see e.g., (Szabó and Fáth 2007; Hofbauer and Sigmund 2003)).

5.2.3 *Classes of Evolutionary Games*

Within this subsection, the possible classes of (2 player)–(2 strategy) games are defined. The first part of this subsection focuses on classes of the symmetric version of the game Γ (see (5.14)), whereas the second part deals with the bimatrix version of the game (see (5.10)).

5.2.3.1 Classes of Symmetric Games

Following the classification scheme of (Weibull 1995) (see also Szabó and Fáth 2007), only three classes of symmetric (2 player)–(2 strategy) games are possible, namely the dominant game class, the class of anti-coordination games, and the coordination game class. For $a < 0$ and $b > 0$ (see (5.14)), the game belongs to the class of dominant games having only one pure NE (s_1^A, s_1^B), which is also the dominant strategy and the only ESS of the game. For $a, b < 0$, the game Γ is an anti-coordination game, having two pure, non-symmetric Nash equilibria $((s_1^A, s_2^B)$ and $(s_2^A, s_1^B))$, and one symmetric interior mixed strategy NE $(\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{b}{a+b}, \frac{b}{a+b})$, which is the only ESS of the game. For $a, b > 0$, the game belongs to the coordination game class, having two pure symmetric Nash equilibria $((s_1^A, s_1^B)$ and $(s_2^A, s_2^B))$, which are the two possible ESSs, and one symmetric interior NE at $(\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{b}{a+b}, \frac{b}{a+b})$. For $b < 0$ and $a > 0$, the game is again a dominant game, having only one pure NE and ESS at (s_2^A, s_2^B).

To illustrate these formal results and visualize the outcomes of the different game classes, this section presents the numerical simulations with different parameter

Table 5.2 Parameter values of the three different sets of symmetric games

Parameter setting	Game class	U_{11}	U_{12}	U_{21}	U_{22}	a	b	Nash equilibria
Set_1	Dominant class	10	4	12	5	−2	1	One pure Nash equilibrium (s_2^A, s_2^B)
Set_2	Coordination class	10	4	9	5	1	1	Two pure Nash equilibria and one interior NE at $s^\star = \frac{1}{2}$
Set_3	Anti-Coord. class	10	7	12	5	−2	−2	Two pure asymmetric Nash equilibria and one interior NE at $s^\star = \frac{1}{2}$

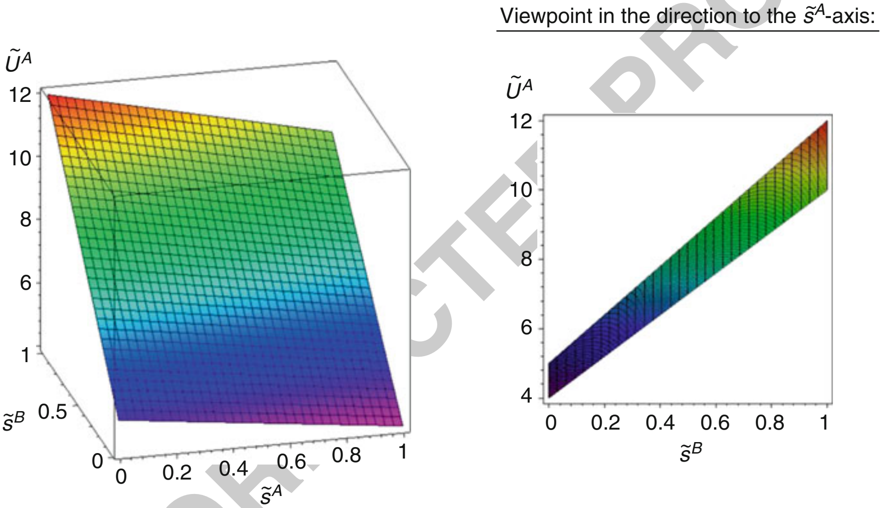


Fig. 5.2 Mixed-strategy payoff function $\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)$ for player A within parameter set Set_1 as a function of the mixed strategies of player A ($\tilde{s}^A$) and B ($\tilde{s}^B$)

settings of symmetric games. The parameter setting Set_1 belongs to the class of dominant games, parameter setting Set_2 belongs to the coordination game class, whereas the setting Set_3 describes an anti-coordination game. Table 5.2 summarizes the different parameters of the three sets.

Dominant Games

Figure 5.2 visualizes the mixed-strategy payoff function $\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)$ (see (5.4)) for player A within parameter set Set_1. The right picture shows a special projection of the surface in which the observer looks in the direction of the $\tilde{s}^A$-axis. The figure

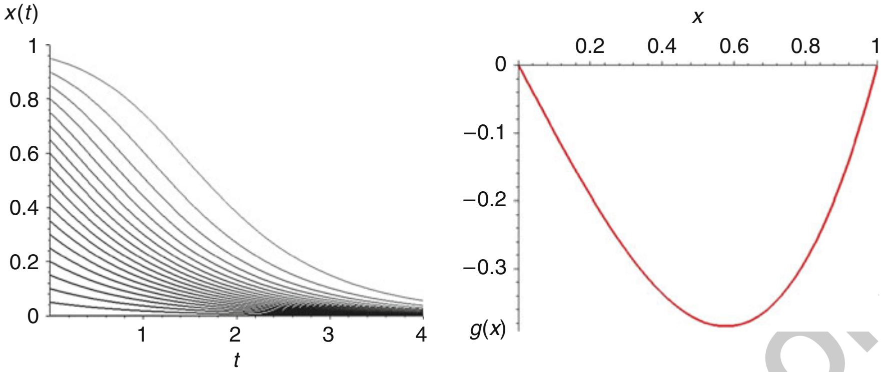


Fig. 5.3 Function $x(t)$, the fraction of players choosing the strategy s_1 at time t, for different initial conditions within parameter set Set_1 (*left picture*). The picture on the right shows the function $g(x)$, which determines the dynamical behavior of $x(t)$

shows that the parameter set Set_1 belongs to the class of dominant games and that only one pure NE exists $((s_2^A, s_2^B)\hat{=}(\tilde{s}^A = 0, \tilde{s}^B = 0))$, which is the dominant strategy of the game. This property can be seen in the left picture of Fig. 5.2 if one fixes the mixed strategy of player B to an arbitrary value $(\tilde{s}^B \in [0, 1])$. The best response for player A will always be the dominant strategy $s_2^A \hat{=} (\tilde{s}^A = 0)$. However, a dilemma appears within Set_1, as the payoff for the dominant strategy combination $(\tilde{\mathcal{U}}^A(\tilde{s}^A = 0, \tilde{s}^B = 0) = 5)$ is far below the highest point of the surface. If both players had chosen the strategy combination $(s_1^A, s_1^B)\hat{=}(\tilde{s}^A = 1, \tilde{s}^B = 1)$, it would have been much better for them $(\tilde{\mathcal{U}}^A(\tilde{s}^A = 1, \tilde{s}^B = 1) = 10)$. The structure of the game within parameter set Set_1 is comparable to a "prisoner's dilemma" game. As no interior NE is present within parameter set Set_1, the partial derivative (see (5.7)) of $\tilde{\mathcal{U}}^A$ does not vanish within the given boundaries. The right picture of Fig. 5.2 visualizes this fact as no cord-up point was found within the special $\tilde{s}^A$-projection.

The right picture of Fig. 5.3 shows the function $g(x)$ within parameter set Set_1, whereas the left picture visualizes the numerical results of replicator dynamics $(x(t)$, see (5.14)) for several initial conditions of the population function $(x(t = 0) = 0, 0.05, 0.1, ..., 0.95)$. As the function $g(x)$ is negative for all $x \in]0, 1[$, the fraction of players choosing the strategy s_1 $(x(t))$ will always decrease until everybody chooses the strategy s_2, independently of the initial condition.

Coordination Games

Within parameter set Set_2, the payoff $U_{21} = 9$ has decreased compared to the value of Set_1 $(U_{21} = 12)$. Due to this decrease, the game class has shifted from the class of dominant games to the coordination game class. The game has now two pure, symmetric Nash equilibria $((s_1^A, s_1^B)\hat{=}(\tilde{s}^A = 1, \tilde{s}^B = 1)$ and $(s_2^A, s_2^B)\hat{=}(\tilde{s}^A = 0, \tilde{s}^B = 0))$ and one interior mixed-strategy Nash equilibrium

this figure will be printed in b/w

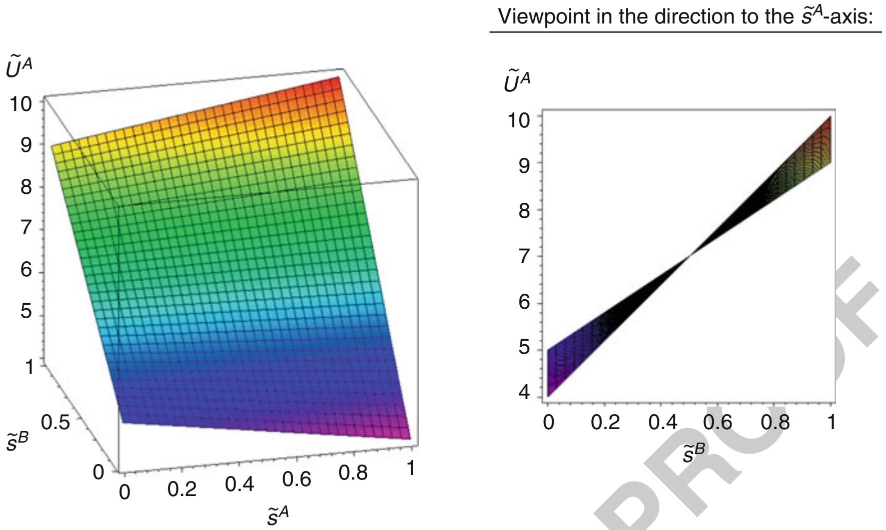


Fig. 5.4 Mixed-strategy payoff function $\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)$ for player A within parameter set Set_2 as a function of the mixed strategies for player A ($\tilde{s}^A$) and B ($\tilde{s}^B$)

$((\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{1}{2}, \frac{1}{2}))$. The apparency of the two pure Nash equilibria is visualized within the left picture of Fig. 5.4. If player B is expected to choose a mixed strategy $\tilde{s}^B > s^\star$, the best response for player A is the pure strategy $s_1 \hat{=} \tilde{s}^A = 1$, whereas if player B is expected to choose a mixed-strategy $\tilde{s}^B < s^\star$, the best response for player A is the pure strategy $s_2 \hat{=} \tilde{s}^A = 0$. The mixed-strategy Nash equilibrium is visualized within the right picture of Fig. 5.4. Due to the fact that the partial derivative of the payoff surface for player A vanishes at the value of the mixed strategy NE, the whole surface shrinks to one point, if one projects the viewpoint in the direction to the $\tilde{s}^A$-axis (see the right picture of Fig. 5.4).

The value of the mixed-strategy Nash equilibrium is equal to the zero point of the function $g(x)$ (see right picture of Fig. 5.5). The function $g(x)$ (which determines the dynamical behavior of the population function $x(t)$) has, beside its negative region ($g(x) < 0 \ \forall \ x \in]0, s^\star[$), also a region where its value is positive ($g(x) > 0 \ \forall \ x \in]s^\star, 1[$). Due to this property, two evolutionary stable strategies emerge ($x(t \to \infty) = 0$ and $x(t \to \infty) = 1$). To which of these ESSs the population will evolve depends on the initial condition. If the fraction of s_1-player types at the initial time $t = 0$ is below the value of the mixed strategy NE ($x(0) < s^\star = 0.5$), the population will evolve to the ESS $\lim\limits_{t \to \infty}(x(t)) = 0$, which corresponds to a population solely choosing the s_2-strategy. Only if the initial fraction is above the mixed strategy threshold ($x(0) > s^\star$), the population will end in the ESS $\lim\limits_{t \to \infty}(x(t)) = 1$. The horizontal population path at $x(0.5) = 0.5$ is an artefact of the numerical simulation and is not an ESS, as the solution is unstable in respect to infinitely small perturbations.

this figure will be printed in b/w

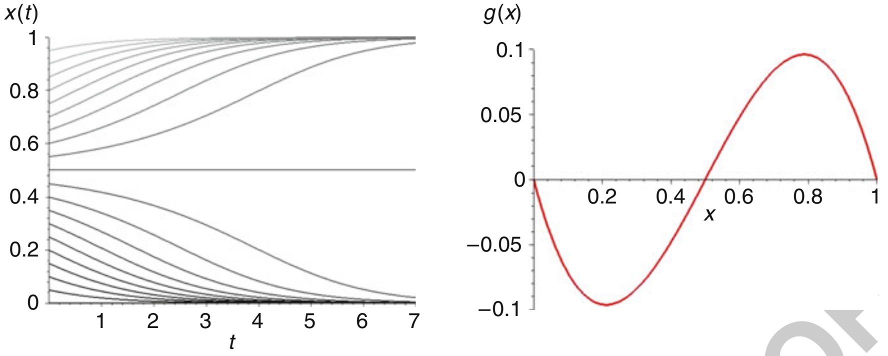


Fig. 5.5 Function $x(t)$, the fraction of players choosing the strategy s_1 at time t, for different initial conditions within parameter set Set_2 (*left picture*). The picture on the right shows the function $g(x)$, which determines the dynamical behavior of $x(t)$

Anti-Coordination Games

Within parameter set Set_3, the payoff $U_{12} = 7$ has increased above the U_{22}-value (Set_3: $U_{22} = 5$). Due to this increase, the game class has shifted towards the class of anti-coordination games. Such games have two asymmetric pure Nash equilibria $((s_1^A, s_2^B)$ and $(s_2^A, s_1^B))$ and one interior mixed-strategy Nash equilibrium, which is the only ESS of such games. The apparency of the two asymmetric Nash equilibria is visualized within the left picture of Fig. 5.6, whereas the mixed-strategy Nash equilibrium (Set_3: $s^\star = 0.5$) is visualized within the right picture.

The value of the mixed-strategy NE is again equal to the zero point of the function $g(x)$ (see right picture of Fig. 5.7). The function $g(x)$ has now a positive region at $(g(x) > 0 \;\forall\; x \;\in]0, s^\star[)$ and a negative region at $(g(x) < 0 \;\forall\; x \;\in]s^\star, 1[)$. Independently of the specific value of the initial condition, the population will always asymptotically end in the ESS $x = s^\star = 0.5$ (see the left picture of Fig. 5.7).

It was shown within this subsection that symmetric (2×2)-games can be separated into three classes. However, if the number of available strategies increases, the number of possible classes also needs to be extended. Zeeman has defined 19 different game classes of symmetric (2×3)-games (Zeeman 1980).

5.2.3.2 Classes of Bimatrix Games

This subsection summarizes the numerical results of the unsymmetric model, where two separate subpopulations play an evolutionary bimatrix game. Following the bimatrix classification scheme of Cressman (2003) (see also

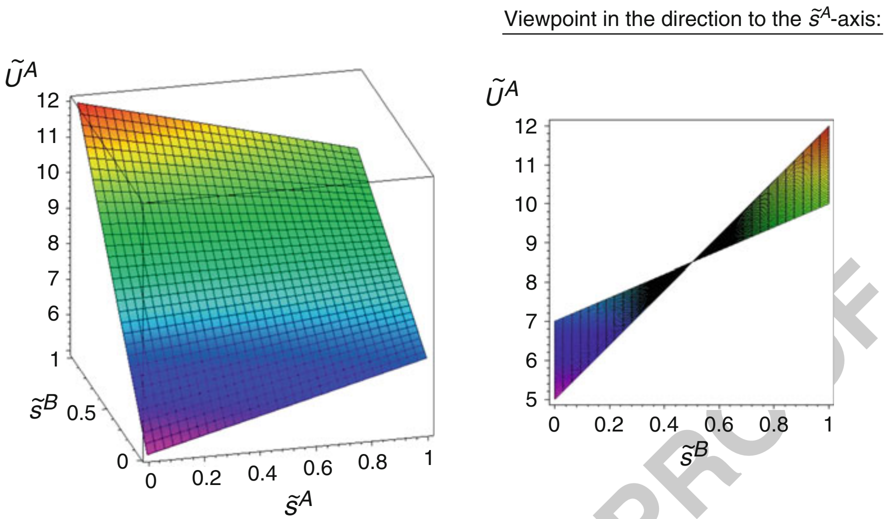


Fig. 5.6 Mixed-strategy payoff function $\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)$ for player A within parameter set Set_3 as a function of the mixed strategies for player A ($\tilde{s}^A$) and B ($\tilde{s}^B$)

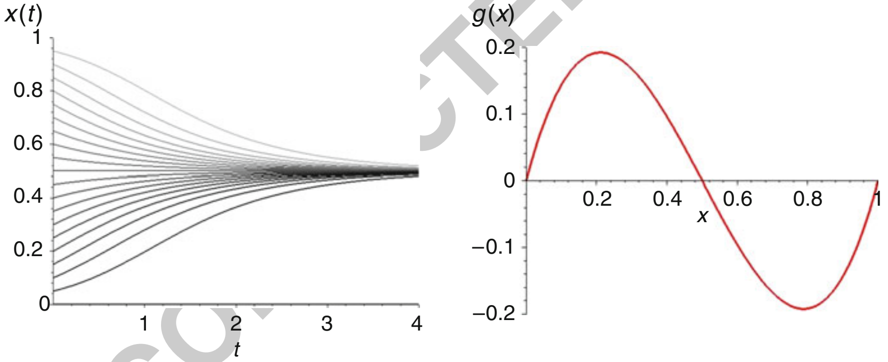


Fig. 5.7 Function $x(t)$, the fraction of players choosing the strategy s_1 at time t, for different initial conditions within parameter set Set_8 (*left picture*). The picture on the right shows the function $g(x)$, which determines the dynamical behavior of $x(t)$

(Szabó and Fáth 2007)), again only three major[3] classes are possible within the unsymmetric version of the game Γ, namely the corner class, the center class and the saddle class. The game belongs to the saddle class if all of the parameters are positive ($a^A, b^A, a^B, b^B > 0$). Saddle-class games have an interior mixed-strategy

---

[3] Beside the three major (generic) classes there exist also degenerate cases, where one ore more of the parameters a^A, b^A, a^B and b^B are zero (see Szabó and Fáth 2007).

Table 5.3 Parameter values of the four different sets of unsymmetric games

Parameter setting	μ	Class of Game μ	U_{11}^{μ}	U_{12}^{μ}	U_{21}^{μ}	U_{22}^{μ}	a^{μ}	b^{μ}	Nash equilibria of game μ	Game class NE and ESS
Set_1^{us}	A:	Dominant class	10	4	14	5	−4	1	One pure NE (s_2^A, s_2^B)	Corner class
	B:	Dominant class	10	12	2	5	−2	3	One pure NE (s_2^A, s_2^B)	One NE being ESS (s_2^A, s_2^B)
Set_2^{us}	A:	Coord. class	10	4	9	5	1	1	Two pure NE, one int. NE $(s^\star = \frac{1}{2})$	Saddle class
	B:	Coord. class	10	7	4	5	3	1	Two pure NE, one int. NE $(s^\star = \frac{1}{4})$	Two ESSs (s_1^A, s_1^B), (s_2^A, s_2^B)
Set_3^{us}	A:	Anti-Co. class	10	7	12	5	−2	−2	Two pure NE, one int. NE $(s^\star = \frac{1}{2})$	Saddle class
	B:	Anti-Co. class	10	12	9	5	−2	−4	Two pure NE, one int. NE $(s^\star = \frac{2}{3})$	Two ESSs (s_1^A, s_2^B), (s_2^A, s_1^B)
Set_4^{us}	A:	Coord. class	10	4	7	5	3	1	Two pure NE, one int. NE $(s^\star = \frac{1}{4})$	Center class
	B:	Anti-Co. class	10	12	9	5	−2	−4	Two pure NE, one int. NE $(s^\star = \frac{2}{3})$	No NE nor ESS

Nash equilibrium at $(\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{b^B}{a^B+b^B}, \frac{b^A}{a^A+b^A})$ and two pure, symmetric Nash equilibria $((s_1^A, s_1^B)$ and $(s_2^A, s_2^B))$, which are the two ESSs of the game. For $a^A, b^A > 0$ and $a^B, b^B < 0$ (or $a^A, b^A < 0$ and $a^B, b^B > 0$), the game describes a center-class game, having only one NE, namely the interior mixed-strategy NE at $(\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{b^B}{a^B+b^B}, \frac{b^A}{a^A+b^A})$. Center-class games do not have any ESS, and the population trajectories are closed cycles. Corner-class games emerge if the parameters fulfill the following conditions: $a^A < 0 < b^A, b^B > 0, a^B \neq 0$ (or $a^B < 0 < b^B, b^A > 0, a^A \neq 0$). Such games have only one pure Nash equilibrium (s_2^A, s_2^B) (or (s_1^A, s_1^B)), which is the dominant strategy and the only ESS of the game.

To illustrate these theoretical results and visualize the outcomes of the different game classes, the parameters were fixed within four different game settings (see Table 5.3). The parameter setting Set_1^{us} belongs to the corner class of bimatrix games, the sets Set_2^{us} and Set_3^{us} are saddle-class games, and the last setting (Set_4^{us}) describes a game that belongs to the center class.

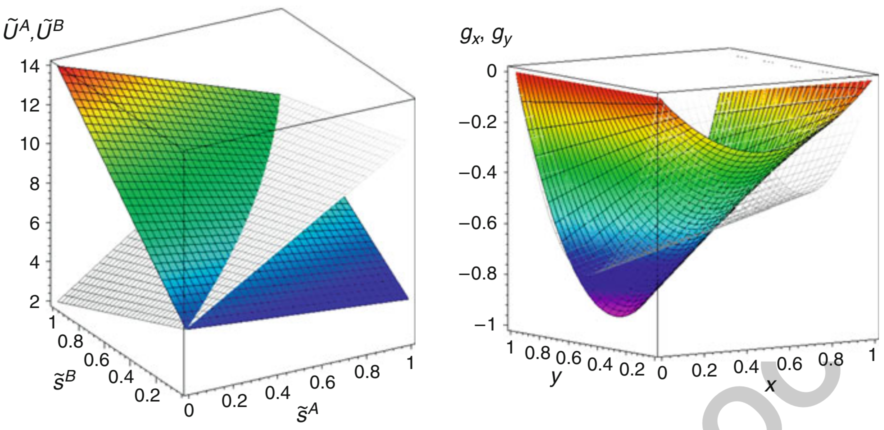


Fig. 5.8 *Left picture*: Mixed-strategy payoff function for player A ($\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)$, *colored surface*) and player B ($\tilde{\mathcal{U}}^B(\tilde{s}^A, \tilde{s}^B)$, *wired grey surface*) within parameter set Set_1^{us} as a function of the mixed strategies of player A ($\tilde{s}^A$) and B ($\tilde{s}^B$). *Right picture*: $g_x(x, y)$ (*colored surface*) and $g_y(x, y)$ (*wired grey surface*) as functions of the strategic population fractions of group A (x) and group B (y)

Corner class

The left picture of Fig. 5.8 visualizes the mixed-strategy payoff function for player A–$\tilde{\mathcal{U}}^A(\tilde{s}^A, \tilde{s}^B)$: colored surface, see (5.4) – and player B – $\tilde{\mathcal{U}}^B(\tilde{s}^A, \tilde{s}^B)$: wired grey surface–within parameter set Set_1^{us}. The set Set_1^{us} is similar to the symmetric parameter set Set_1 of a prisoner's dilemma game. In contrast to the set Set_1, the two game matrices for player A and B are unsymmetric ($U_{12}^A = 4 \neq 2 = U_{21}^B$ and $U_{21}^A = 14 \neq 12 = U_{12}^B$). The structure of the surfaces indicates that both groups have again only one NE, which is the dominant strategy $(s_2^A, s_2^B) \hat{=} (\tilde{s}^{A*} = 0, \tilde{s}^{B*} = 0)$.

The right picture of Fig. 5.8 displays the two functions $g_x(x, y)$ (colored surface) and $g_y(x, y)$ (wired grey surface) that determine the dynamical behavior of the strategical decisions of group A ($x(t)$) and group B ($y(t)$) (see (5.10)). The amount of players choosing strategy s_1 will in both groups monotonically decrease and will – independently of the initial value – finally reach the only ESS ($x = 0$, $y = 0$), because the two surfaces are always below or equal to zero ($g_x(x, y) \leq 0$, $g_y(x, y) \leq 0 \; \forall \; x, y \in [0, 1]$).

The evolution of the strategic behavior of the two groups is visualized in Fig. 5.9. The plot describes the numerical results of (5.10) for three different initial conditions, displayed through the three colored curves (xy-trajectories). The three trajectories are embedded in a field-plot phase diagram, where the little grey arrows describe the direction of a "strategic wind" the population has to follow during its time evolution. The three initial conditions ($x(0), y(0)$) are marked with colored circles at the beginning of the three curves. The several colored arrows which are on top of the trajectories describe the population movement for some intermediate time steps, where the length of arrows indicate the absolute value of the strategic change

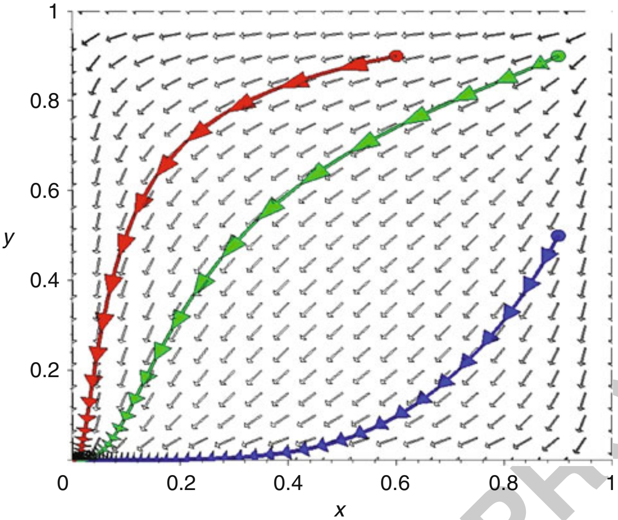


Fig. 5.9 Phase diagram of the xy-trajectories for three different initial conditions within parameter set Set_1^{us}. x describes the fraction of players within group A choosing the strategy s_1, while y is a similar fraction within group B

velocity within the population. Within Fig. 5.9, the difference in the intermediate time steps ($\delta t = 0.125$) is equal for all three trajectories. The unsymmetric behavior of the trajectories is due to the unsymmetry of the parameter set. The green curve, for example, starts at a symmetric initial value ($x(0) = 0.9, y(0) = 0.9$), but as time evolves, it follows an unsymmetric evolution.

The interpretation of the results of Fig. 5.9 is comparable to the results for the parameter set Set_1 of the symmetric model. Both population subgroups play a prisoner's dilemma game and the evolution of their strategical choice will finally – independently of the initial condition – reach a state where everybody chooses the dominant strategy s_2. Similar to the symmetric model, the players face a dilemma, as the two populations evolve towards a low-payoff ESS ($\tilde{\mathcal{U}}^\mu(0,0) = 5 < 10 = \tilde{\mathcal{U}}^\mu(1,1)$). The game category belongs formally to the corner class. The velocity of the strategic change (length of the colored arrows) of the three trajectories differs slightly during the evolution. In the middle region of the trajectories, the velocity is the highest, whereas at the end (near to the ESS), the strategic change slows down very much.

Saddle class

Within the parameter set Set_2^{us}, both subpopulations play a coordination game. A bimatrix game that is composed of two coordination games always results in a saddle-class game. The structure of the payoff surfaces (see left picture in Fig. 5.10)

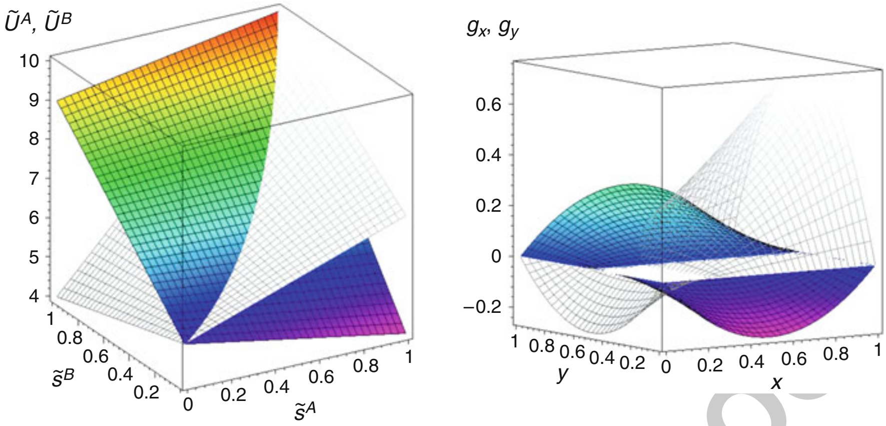


Fig. 5.10 Payoffs and functions $g_x(x, y)$ and $g_y(x, y)$ within set Set_2^{us}; similar to the description in Fig. 5.8

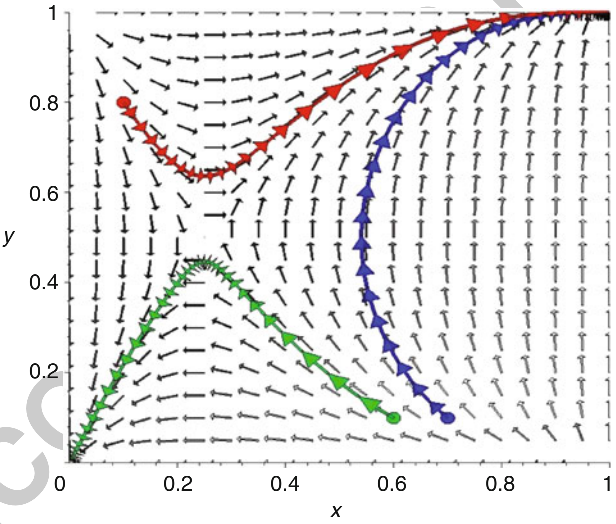


Fig. 5.11 Phase diagram for three different xy-trajectories within set Set_2^{us}; similar to the description in Fig. 5.11

indicates that both groups have now two pure Nash equilibria ((s_1^A, s_1^B) and (s_2^A, s_2^B)). Additionally, there exists an interior mixed strategy NE ($(\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{1}{2}, \frac{1}{4})$). To indicate the zero-level, an additional white plane was added to Fig. 5.10 (right hand side). Within this parameter set, the two surfaces have regions where they have positive values ($g_x(x, y) > 0 \ \forall \ y \in]\tilde{s}^{B\star}, 1]$ and $g_y(x, y) > 0 \ \forall \ x \in]\tilde{s}^{A\star}, 1]$) and regions where they are negative ($g_x(x, y) < 0 \ \forall \ y \in]0, \tilde{s}^{B\star}[$ and $g_y(x, y) > 0 \ \forall \ x \in]0, \tilde{s}^{A\star}[$). The interior mixed strategy NE is exactly at the point

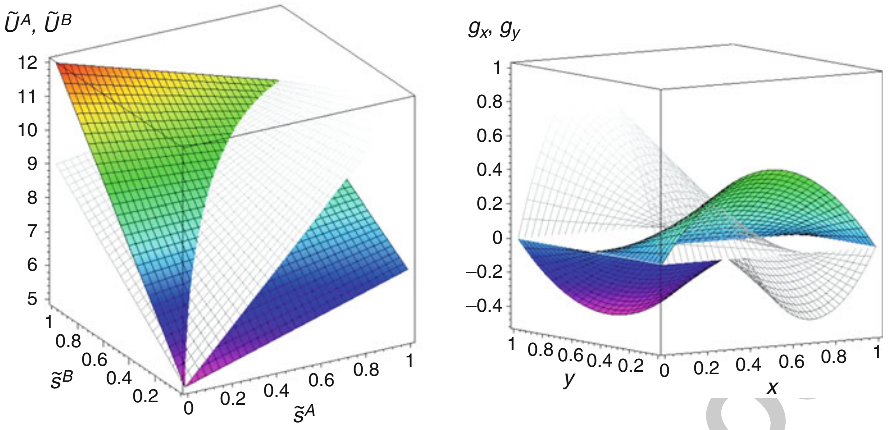


Fig. 5.12 Payoffs and functions $g_x(x, y)$ and $g_y(x, y)$ within set Set_3^{us}; similar to the description in Fig. 5.8

where all of the three surfaces intersect. As all of the parameters (a^A, a^B, b^A, b^B) are positive, the game category belongs to the saddle class of bimatrix games and it has two symmetric ESSs.

The results of the evolutionary game of parameter set Set_2^{us} are visualized in Fig. 5.11. As the strategic change velocities of the three different trajectories are quite different, the time steps (δt) between the colored arrows are not the same for the three different population paths. The red and green trajectories have the same time increment $(\delta t = 0.35)$, whereas the arrows on the blue path are separated by a time lag of $\delta t = 2$. The strategic change of the blue population path is the slowest; starting from an initial condition $(x(0) = 0.7, y(0) = 0.1)$, the fraction of players who choose the s_1-strategy monotonically decreases within group B $(y(t))$, while within group A $(x(t))$, the s_1-fraction first decreases and then increases until the whole population finally ends in the ESS $(s_1^A, s_1^B) \hat{=} (\tilde{s}^{A*} = 1, \tilde{s}^{B*} = 1)$ (all players choose the s_1-strategy). The red trajectory, which starts at the initial condition $(x(0) = 0.1, y(0) = 0.8)$, also ends within the ESS (s_1^A, s_1^B). Its strategic change velocity, however, slows down very much at the region near the interior NE. The initial condition of the green trajectory $(x(0) = 0.6, y(0) = 0.1)$ is only slightly different from the initial value of the blue curve; its evolution, however, is totally different. The s_1-fraction monotonically decreases within group A $(x(t))$, while within group B $(y(t))$, the s_1 fraction first increases and then decreases, until the whole population finally ends in the ESS $(s_2^A, s_2^B) \hat{=} (\tilde{s}^{A*} = 0, \tilde{s}^{B*} = 0)$ (all players choose the s_2-strategy). Similar to the red curve, the strategic change velocity slows down very much at the region near to the interior NE.

Parameter set Set_3^{us} is a saddle-class bimatrix game in which both subpopulations play an anti-coordination game. The structure of the payoff surfaces (see left picture in Fig. 5.12) indicates that both groups have two asymmetric pure Nash equilibria

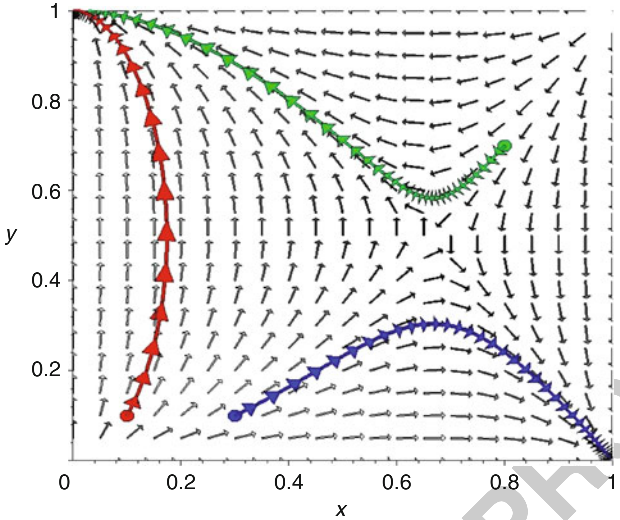


Fig. 5.13 Phase diagram for three different xy-trajectories within set Set_3^{us}; similar to the description in Fig. 5.9

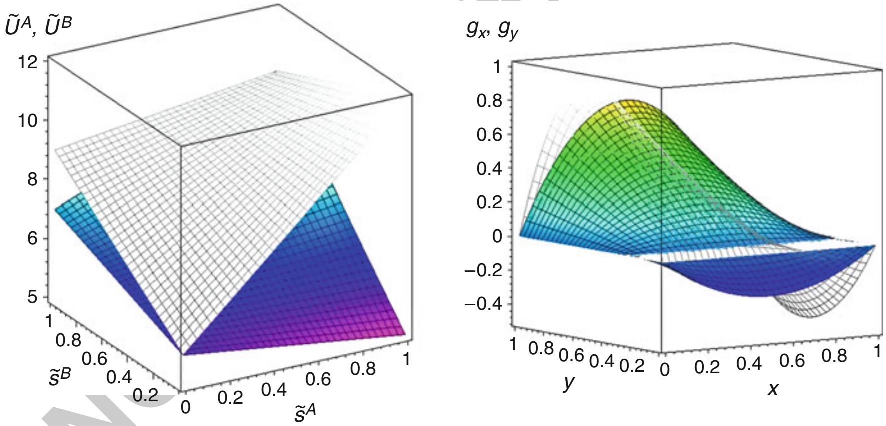


Fig. 5.14 Payoffs and functions $g_x(x, y)$ and $g_y(x, y)$ within set Set_4^{us}; similar to the description in Fig. 5.8

$((s_1^A, s_2^B)$ and $(s_2^A, s_1^B))$ and one interior mixed strategy NE $((\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{1}{2}, \frac{2}{3}))$. As all of the parameters (a^A, a^B, b^A, b^B) are negative, the game category belongs to the saddle class of bimatrix games, and it has two asymmetric ESSs.

The results of the evolutionary game of parameter set Set_3^{us} are visualized in Figs. 5.12 and 5.13. The time steps (δt) between the colored arrows are the same for all three population paths $(\delta t = 0.125)$.

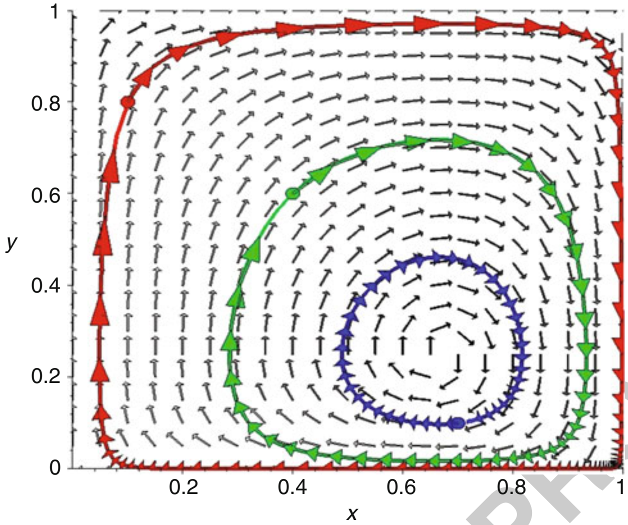


Fig. 5.15 Phase diagram for three different xy-trajectories within set Set_4^{us}; similar to the description in Fig. 5.9

Center class

Finally, the last parameter set (Set_4^{us}) belongs to the category of center-class games. Within parameter set Set_4^{us}, the subpopulation A plays a coordination game, while subpopulation B plays an anti-coordination game. The structure of the payoff surfaces (see left picture in Fig. 5.14) indicates that there is only one interior mixed-strategy NE $((\tilde{s}^{A\star}, \tilde{s}^{B\star}) = (\frac{1}{4}, \frac{2}{3}))$.

The results of the evolutionary game of parameter set Set_4^{us} are visualized in Fig. 5.15 and show that all of the trajectories cycle around the interior NE, which indicates the absence of an ESS. The time needed for one cycle is larger for bigger cycles and, as a result, the time steps (δt) between the colored arrows are the smallest for the blue trajectory $(\delta t = 6.5)$ and the biggest for the red closed curve $(\delta t = 14.5)$ (green: $\delta t = 8$).

5.3 Applications

In recent years, the market of scientific publishing faces several forces that may cause a major change of traditional market mechanisms. Currently, two main approaches have emerged. On the one hand, new open-access journals are brought to being, either through transformation of traditional journals or through creation of new titles. This approach is often called the “Golden Road to Open Access.” On the other hand, authors may self-archive their articles in institutional or subject-

Table 5.4 Researchers' open-access payoff matrix

A\B	o	ø
o	$(r+\delta, r+\delta)$	$(r-\alpha, r+\beta)$
ø	$(r+\beta, r-\alpha)$	(r, r)

based repositories, a model referred to as the "Green Road to Open Access" (Harnad 2005; Guedon 2004). The digital revolution of the information age and, in particular, the sweeping changes of scientific communication brought about by computing and novel communication technology, potentiate global, high-grade scientific information for free. The arXiv, for example, is the leading scientific communication platform, mainly for mathematics and physics, to which everyone in the world has free access on. In the following, we understand open-access publishing as the electronic publication of scientific information on a platform that provides access to this information for all potential users, without financial or other barriers. In contrast, most other scientific disciplines do not make use of open-access publishing, even though they support this model if asked for (Deutsche Forschungsgemeinschaft 2006; Schroter et al. 2005). Instead, they submit research papers to traditional journals that do not provide free access to their articles. Considering that the majority of scientists regard open-access publishing as superior to the traditional system, one may question why it is adopted only by a few disciplines.

5.3.1 *Scientific Communication and the Open-Access Decision*

Based on the assumption that the main goal of scientists is the maximization of their reputation, we try to answer this question from the perspective of the producers of scientific information by using a game-theoretical approach. Scientific reputation originates mainly from two different sources: on the one hand, the citations to the articles of a scientist, and on the other hand, the reputation of the journals in which she/he publishes her/his articles (Dewett and Denisi 2004). Starting from a general symmetric (2 player)–(2 strategy) game Γ (see definition (5.3)), where two authors have to decide whether they publish open access or not, different possible game settings are developed. This application focuses on a one-population model of an open-access game of scientific communication and extends it to an evolutionary game (for details, see Hanauske et al. 2007, 2010b).

To describe the underlying open-access game, we use a normal-form representation of a two-player game Γ where each player (Player 1 $\hat{=}$ A, Player 2 $\hat{=}$ B) can choose between two strategies ($\mathcal{S}^A = \{s_1^A, s_2^A\}$, $\mathcal{S}^B = \{s_1^B, s_2^B\}$). In our case, the two strategies represent the authors' choice between publishing open access (o) or not (ø). The whole strategy space $\mathcal{S}$ is composed with use of a Cartesian product of the individual strategies of the two players (scientists):

$$\mathcal{S} = \mathcal{S}^A \times \mathcal{S}^B = \{(\mathrm{o},\mathrm{o}), (\mathrm{o},\text{ø}), (\text{ø},\mathrm{o}), (\text{ø},\text{ø})\} \qquad (5.15)$$

As outlined before, it is assumed that the main objective of scientists is the maximization of their reputation. In the following, we focus on a situation where the two scientists belong to a scientific community in which the open-access paradigm is not yet broadly adopted, and the publishers decline the acceptance of articles that are already accessible on an open access server. The payoff structure of this game is modeled by the following payoff matrix (Table 5.4).

AQ2

The actual reputation of the two scientists is represented by a single parameter r.[4] If both players decide to publish their papers only in traditional journals (ø,ø), their reputation r does not change. If only one of the two players chooses the open-access strategy ((ø,o) or (o,ø)), the parameters α and β ($\alpha, \beta \geq 0$) describe the decrease and the increase of the scientists' reputation, depending on the selected strategy. By modeling the payoff in this way, it is assumed that the reputation of the player who performs open access decreases if the other player simultaneously decides not to publish open access. This can be explained by the fact that in "non-open-access communities," reputation is mainly defined through the reputation of the journals in which a scientist publishes. Thus, if performing open-access (making publication in traditional journals impossible), the scientist has no chance to gain journal-related reputation anymore. On the other hand, the parameter β describes the potential increase of reputation of a scientist who refuses to perform open-access, while the other player selects the open-access strategy. The parameter δ represents the potential benefit in the case that both players choose the open-access strategy (o,o). The payoff for each player then is $r + \delta$. In this case, it is assumed that if both players choose the open-access strategy, the publishers are forced to accept articles for publication even if they are already accessible (see also the application discussed in Sect. 5.3.2). Then, scientists can gain reputation both through the reputation of the journal they publish in and through the increase of citations due to a broader accessibility (Lawrence 2001; Harnad and Brody 2004; Eysenbach 2006).

As the presented open-access game is a symmetric game and the parameter $b = \alpha$ is positive, the underlying game class depends only on the sign of the parameter $a = \delta - \beta$. For $\delta > \beta$, the game belongs to the class of coordination games, whereas for $\delta < \beta$, the game has the structure of a dominant game with a dilemma. For example, if the payoff parameters are fixed to the values $\alpha = 1$, $\beta = 2.25$, and $\delta = 0.25$ ($a = -2$ and $b = 1$), the results of the open-access game would be identical to the parameter setting Set_1 of the dominant game presented in Sect. 5.2.3.1. Although the payoff for both players would be higher if they chose the strategy set (o,o), they are stuck within the Nash equilibrium (ø,ø). This outcome describes the paradox situation of many scientific disciplines: On the one hand, scientists realize that they would benefit if all players adopt open access, but on the other hand, no player has an individual incentive to change. For $\alpha = 1$, $\beta = 0.25$, and $\delta = 1.25$ ($a = 1$ and $b = 1$), the game belongs to the class of coordination games, and its corresponding results are also discussed in Sect. 5.2.3.1

---

[4] By using this formalization, we assume that both scientists are on a similar level of reputation. If they would have different "starting" reputation values, the game would be unsymmetric.

Table 5.5 Payoff matrix of the "Author(A)–Journal(B)" open-access bimatrix game

A\B	o	ø
o	$(r+\delta+I\ ,\ r-\kappa)$	$(r+\delta\ ,\ 0)$
ø	$(r+I\ ,\ r)$	$(r-P+I\ ,\ r+P)$

(see parameter setting Set_2). In contrast to set Set_1, this game has two pure Nash equilibria ((o,o) and (ø,ø)) and one mixed-strategy Nash equilibrium $\frac{1}{2}$(o,o). (o,o) is payoff dominant, whereas (ø,ø) is the risk-dominant pure Nash equilibrium. The mixed-strategy Nash equilibrium $\frac{1}{2}$(o,o) implies that one scientist has the incentive to choose non-open-access if she/he expects the probability of the other player to choose non-open-access to be higher than 50 % (for further details see (Hanauske et al. 2007)). As $b = \alpha > 0$, the class of the open-access game cannot be parameterized as an anti-coordination game.

5.3.2 *Evolution of Hub-and-Spoke Communication Networks*

Within this subsection, the interconnected network of scientific journals and researchers is modeled as an unsymmetric bimatrix game. This application is an example of a more general analysis of a "Hub-and-Spoke Communication Network," which is currently under investigation (Hanauske et al. 2010a). The main actors within the scientific communication network are the authors of scientific articles (Spokes, population group A) and the scientific journals (Hubs, population group B). Following the approach of Habermann (Habermann and Habermann 2009), but restricting the focus to green open access, the researchers have two possible strategies $\{s_1^A, s_2^A\} = \{$o,ø$\} \hat{=} \{$*publishing open access*, *conventional publishing*$\}$. Within the underlying game, the group of scientific journals have the following two strategies: $\{s_1^B, s_2^B\} = \{$o,ø$\} \hat{=} \{$*accept open access*, *decline open access*$\}$. Table 5.5 describes one possible way of a parameterization of the "Author(A)–Journal(B)" open-access bimatrix game (see also (Habermann and Habermann 2009) for another kind of parameterization). Similar to what was introduced in Sect. 5.3.2, the parameter r describes the reputation of the scientist and the parameter δ quantifies the author's potential benefit if she/he chooses the open-access strategy o. The parameter I describes the author's additional increase in reputation if she/he publishes her/his new article within the journal (e.g., the journal's impact factor). Parameter κ is meant as a quantity that measures the journal's hypothetical payoff decrease due to fears of a totally green-open-access publishing market. Finally, the parameter P quantifies the possibility of an extraordinary journal price increase due to the journal's market power in a totally conventional publishing market. Taking the parameterization of Table 5.5, the underlying class is only dependent on the following parameters: $a^A = \delta$, $b^A = I - P - \delta$, $a^B = r - \kappa$, and $b^B = P$. Because $a^A = \delta > 0$ and $b^B = P > 0$, the game category cannot belong to the center-class

this figure will be printed in b/w

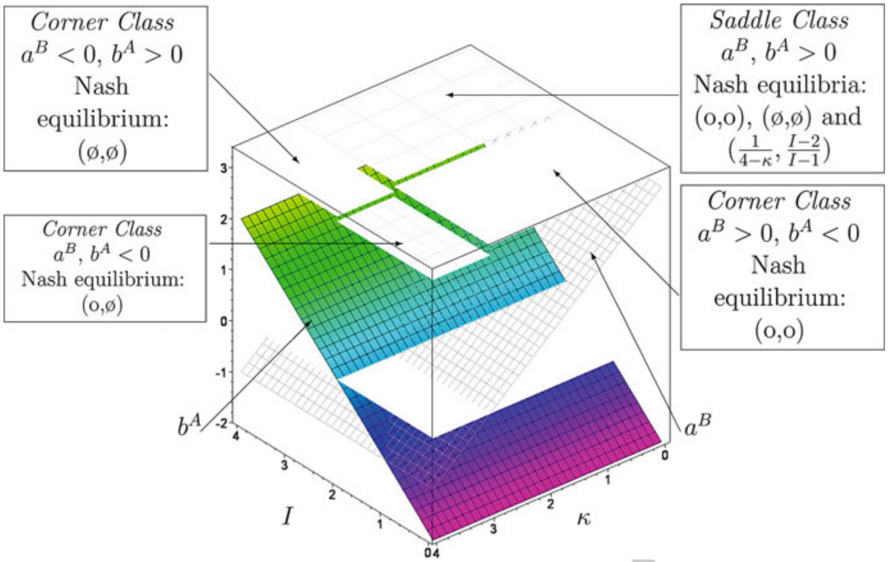


Fig. 5.16 $b^A = I - 2$ (*solid, colored surface*) and $a^B = 3 - \kappa$ (*wired surface*) as a function of the parameters I and κ. The other parameters are fixed to the values: $\delta = 1$, $r = 3$ and $P = 1$

games.[5] For $b^A, a^B > 0$ $(r > \kappa, I > P + \delta)$, the game's category belongs to the saddle-class having two pure, symmetric Nash equilibria $((s_1^A, s_1^B)\hat{=}$(o,o) and $(s_2^A, s_2^B)\hat{=}$(ø,ø)) and one mixed strategy NE at $((\tilde{s}^{A\star}, \tilde{s}^{B\star}) = \left(\frac{P}{r-\kappa+P}, \frac{I-P-\delta}{I-P}\right))$. The outcome of such a parameterization is comparable to the results discussed in Sect. 5.2.3.2 (parametration set Set_2^{us}). For all other parameterizations, the category of the author-journal open-access game belongs to the corner class. For ($b^A < 0$ and $a^B > 0$), the only NE is (o,o), for ($b^A > 0$ and $a^B < 0$), the only NE is (ø,ø), and finally for ($a^B, b^A < 0$), there exists only the asymmetric NE (o,ø).

To visualize these outcomes, Fig. 5.16 shows the different possible classes within the author-journal open-access game for a certain parameterization. The solid, colored surface depicts the parameter b^A as a function of the two payoff parameters κ and I (the other parameters were fixed to the following values: $\delta = 1$, $r = 3$ and $P = 1$). The wired grey surface depicts the parameter a^B, and the solid white surface indicates the zero level. The point where all of the three surfaces intersect $(b^A(\kappa^\circ, I^\circ) = a^B(\kappa^\circ, I^\circ) = 0 \;\rightarrow \kappa^\circ = 3\,,\; I^\circ = 2)$ defines the class boundary. Only for $\kappa > \kappa^\circ$, $I > I^\circ$ is a saddle-class game is realized, whereas in all of the other parameterizations, only one NE and ESS is possible, as the game belongs under such parametrisations to the corner class (for details see (Hanauske et al. 2010b)).

---

[5] Other parameterizations do, however, result in open-access center-class games (Habermann and Habermann 2009).

5.4 Summary and Outlook

One of the main criticism of EGT is the fact that the theory is based on a totally connected network of an infinitely large number of actors, where every player (in each time interval) chooses her/his game partner randomly. In reality, the players are often organized in groups, and even within these groups the players often are not fully connected to all of the group members. The theory of social grouping in decision-based interacting complex networks is one of the most interesting topics within the presented research field. *Evolutionary Game Theory on Complex Networks* is a more realistic framework to simulate population dynamics; however, it often needs a variety of additional parameters to classify the network topologies and updating rules (see e.g., (Szabó and Fáth 2007; Miekisz 2008)).

A second, more recently developed model that tries to implement social grouping into classical[6] evolutionary game theory is *Evolutionary Quantum Game Theory*. Quantum game theory is a mathematical and conceptual amplification of classical game theory. The space of all conceivable decision paths is extended from the purely rational, measurable space in the Hilbertspace of complex numbers. Through the concept of a potential entanglement of the imaginary quantum strategy parts, it is possible to include corporate decision paths, caused by cultural or moral group standards. In quantum game theory, players may cooperate, depending on the degree of entanglement γ among players. The notion of entanglement is perhaps most clearly expressed in terms of Adam Smith's classical concept of sympathy or "fellow feeling," which is a cornerstone of Smith's understanding of individual behavior (Hanauske and Schäfer 2009). In his "*Theory of Moral Sentiments*" (1759) (Sugden 2002), Smith claims that there is a general tendency for fellow-feeling among human beings, whereas the greater the strength of fellow-feeling is, the more closely related the individuals are. For example, there tends to be more fellow-feeling between friends than between acquaintances, and more between close relatives than between distant ones. Fellow-feeling as the human capacity to emphasize and become entangled with others is inversely related to the perceived and felt distance, whereas distance has been interpreted in terms of psychological and physical distance (Sally 2001). It can be shown that Emma and Hans are able to escape the dilemma if their strength of fellow-feeling (strength of strategic entanglement) is high enough to overcome the game's γ-threshold. If this strategy entanglement is large enough, then additional Nash equilibria can occur, previously present dominant strategies could become nonexistent, and new evolutionary stable strategies might appear (see e.g., (Hanauske 2011)).

Within this chapter, the framework of classical EGT has been described in detail. After a general introduction and a brief literature review, the groundings of EGT (Sect. 5.2) have been explained in detail. The formal mathematical model,

---

[6] Following the scientific classification of the physical literature, the notation "classical" is used to describe the scientific sub-discipline that do not use "quantum" concepts to describe the underlying natural processes (example in physics: *Classical Mechanics* vs. *Quantum Mechanics*).

the different concepts of equilibria, and the various classes of evolutionary games have been defined, explained, and visualized to understand the main ideas of EGT. Additionally, in Sect. 5.3 two applications have been discussed:

- Application 1: Scientific communication and the open-access decision (see Sect. 5.3.1)
- Application 2: Evolution of Hub-and-Spoke Communication Networks (see Sect. 5.3.2)

Key points By analysing the game structure of a specific decision problem, policy-makers can learn a lot about the problems they attempt to address. To analyse the problem game theoretically, you need only three things:

- Who is playing the game? Definition of the set of players.
- What can the players do? Definition of the set of actions (strategies) available for each player.
- How much can the players win or lose? Definition of the payoff structure of the underlying game.

If the decision problem can be modelled as a symmetric (two player)–(two strategy) game and you know the payoff structure (define the parameters U_{11}, U_{12}, U_{21} and U_{22} and calculate $a := U_{11} - U_{21}$ and $b := U_{22} - U_{12}$), your game belongs to the following class:

- $b < 0$ and $a > 0$ (or $b > 0$ and $a < 0$): Dominant class
- $a, b > 0$: Coordination class
- $a, b < 0$: Anti-coordination class

If your game belongs to the dominant class and there is no dilemma, use the dominant strategy. If your game belongs to the dominant class and there is a dilemma (or it belongs to the coordination class with a high and low Nash equilibrium, or to the anti-coordination class with a dilemma), you have to think about how much fellow-feeling you have with your game partner – perhaps your socio-economic system is strong enough to escape the game's dilemma.

References

Amann E (1999) Evolutionäre Spieltheorie: Grundlagen und neue Ansätze. Studies in Contemporary Economics. Physica-Verlag, Heidelberg

Axelrod R (1997) The complexity of cooperation: Agent-based models of conflict and cooperation. Princeton Studies in Complexity. Princeton University Press, Princeton, NJ

Bernius S, Hanauske M (2007) Open access. Wirtschaftsinformatik 49(6):456–459 (DOI: 10.1007/s11576-007-0098-2)

Bernius S, Hanauske M, Dugall B, König W (2010) Exploring the effects of a transition to open access: Insights from a simulation study. Unpublished manuscript, submitted to the Journal of the American Society for Information Science and Technology (JASIST)

Clemens C, Riechmann T (2006) Evolutionary dynamics in public good games. Comput Econ 28(4):399–420 (DOI: 10.1007/s10614-006-9044-4)

Cressman R (2003) Evolutionary dynamics and extensive form games. Economic Learning and Social Evolution, vol 5. MIT Press, Cambridge, MA

Deutsche Forschungsgemeinschaft (2006) Publication strategies in transformation? Results of a study on publishing habits and information acquisition with regard to open access. Wiley-VCH Verlag, Weinheim, available online at the URL: http://www.dfg.de/download/pdf/dfg_im_profil/evaluation_statistik/programm_evaluation/studie_publikationsstrategien_bericht_en.pdf

Dewett T, Denisi AS (2004) Exploring scholarly reputation: It's more than just productivity. Scientometrics 60(2):249–272 (DOI: 10.1023/B:SCIE.0000027796.55585.61)

Enquist M, Ghirlanda S (2007) Evolution of social learning does not explain the origin of human cumulative culture. J Theor Biol 246(1):129–135 (DOI: 10.1016/j.jtbi.2006.12.022)

Eysenbach (2006) Citation advantage of open access articles. PLoS Biol 4(5):e157 (DOI: 10.1371/journal.pbio.0040157)

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296(5568):750–752 (DOI: 10.1126/science.1068696)

Guédon JC (2004) The "green" and "gold" roads to open access: The case for mixing and matching. Serials Rev 30(4):315–328 (DOI: 10.1016/j.serrev.2004.09.005), also available online at the URL: http://hdl.handle.net/10760/5860

Habermann K, Habermann L (2009) An evolutionary game-theoretic approach to open access. ArXiv preprint arXiv:0903.4562

Hanauske M (2009) Neue Entwicklungen in der Evolutionären Spieltheorie. Lecture notes of lectures given at the Université Lumière Lyon 2 in Lyon, France (MINERVE Exchange Program). Available online at the URL: http://evolution.wiwi.uni-frankfurt.de/Lyon2009/

Hanauske M (2011) Evolutionäre Quanten-Spieltheorie im Kontext sozio-ökonomischer Systeme/Evolutionary quantum game theory in the context of socio-economic systems. PhD-Thesis. Goethe-Universität Frankfurt am Main, Frankfurt am Main

Hanauske M, Bernius S, Dugall B (2007) Quantum game theory and open access publishing. Physica A 382(2):650–664 (DOI: 10.1016/j.physa.2007.04.012), also available as arXiv preprint arXiv:physics/0612234

Hanauske M, Huber T, Kude T (2010a) *Evolutionary quantum game theory and hubs- and spoke-networks*. Manuscript in preparation

Hanauske M, König W, Dugall B (2010b) Evolutionary quantum game theory and scientific communication. Accepted article at the "Second Brasilian Workshop of the Game Theory Society in honor of JOHN NASH, on the occasion of the 60th anniversary of Nash equilibrium, São Paulo – SP – Brazil July 29 to August 4, 2010", available online at the URL: http://evolution.wiwi.uni-frankfurt.de/BWGT2010/

Hanauske M, Kunz J, Bernius S, König W (2009) Doves and hawks in economics revisited: An evolutionary quantum game theory-based analysis of financial crises. Physica A 389(21):5084–5102 (DOI: 10.1016/j.physa.2010.06.007), also available as arXiv preprint arXiv:physics/0612234

Hanauske M, Schäfer S (2009) Fellow-feeling and cooperation (A quantum game theory-based analysis of a prisoner's dilemma experiment). Unpublished manuscript, originating from the PhD-seminar "Experimental Economics" organised by Prof. Kosfeld, (available from the author)

Harms W, Skyrms B (2008) Evolution of moral norms. In: Ruse M (ed) Oxford handbook on the philosophy of biology. Oxford University Press, Oxford, Chap 18, pp 434–450 (DOI: 10.1093/oxfordhb/9780195182057.003.0019)

Harnad S, Brody T (2004) Comparing the impact of open access (OA) vs. non-OA articles in the same journals. D-Lib Magazine 10(6) (DOI: 10.1045/june2004-harnad)

Harnad S (2005) Fast-forward on the green road to open access: The case against mixing up green and gold. Ariadne (42), available online at the URL: http://www.ariadne.ac.uk/issue42/harnad/intro.html

Hofbauer J, Sigmund K (2003) Evolutionary game dynamics. Bull Am Math Soc 40(4):479–519 (DOI: 10.1090/S0273-0979-03-00988-1)

Janssen M, Ostrom E (2006) Governing social-ecological systems. In: Tesfatsion L, Judd KL (eds) Handbook of computational economics, Volume 2: Agent-based computational economics. Handbooks in Economics, vol 13. North-Holland, Amsterdam, Chap 30, pp 1465–1509 (DOI: 10.1016/S1574-0021(05)02030-7)

Kerr B, Riley MA, Feldman MW, Bohannan BJM (2002) Local dispersal promotes biodiversity in a real-life game of rock paper scissors. Nature 418(6894):171–174 (DOI: 10.1038/nature00823)

Lawrence S (2001) Free online availability substantially increases a paper's impact. Nature 411(6837):521 (DOI: 10.1038/35079151)

Miekisz J (2008) Evolutionary game theory and population dynamics. In: Capasso V, Lachowicz M (eds) Multiscale problems in the life sciences: From microscopic to macroscopic: Lectures from the Banach Center and C.I.M.E. Joint Summer School held in Będlewo, September 4–9, 2006. Lecture Notes in Mathematics, vol 1940. Springer-Verlag, Berlin; Fondazione C.I.M.E., Florence, pp 269–316 (DOI: 10.1007/978-3-540-78362-6_5), also available as arXiv preprint arXiv:q-bio/0703062

Nowak MA, Sigmund K (2002) Bacterial game dynamics. Nature 418(6894):138–139 (DOI: 10.1038/418138a)

Nowak MA, Sigmund K (2003) Evolutionary dynamics of biological games. Science 303(5659):793–799 (DOI: 10.1126/science.1093411), also available online as International Institute for Applied Systems Analysis (IIASA) Laxenburg interim report IR-04-013 at the URL http://www.iiasa.ac.at/Admin/PUB/Documents/IR-04-013.pdf

Ostrom E (2000) Collective action and the evolution of social norms. J Econ Perspect 14(3):137–158 (DOI:10.1257/jep.14.3.137, stable JSTOR URL: http://www.jstor.org/stable/2646923)

Ostrom E (2009) A general framework for analyzing sustainability of social-ecological systems. Science 325(5939):419–422 (DOI: 10.1126/science.1172133)

Pawlowitsch C (2007) Finite populations choose an optimal language. J Theor Biol 249(3):606–616 (DOI: 10.1016/j.jtbi.2007.08.009)

Sally D (2001) On sympathy and games. J Econ Behav Organ 44(1):1–30 (DOI:10.1016/S0167-2681(00)00153-0)

Sandholm WH (2010) Population games and evolutionary dynamics. Economic Learning and Social Evolution. MIT Press, Cambridge, MA

Schlee W (2004) Einführung in die Spieltheorie: Mit Beispielen und Aufgaben. Fried. Vieweg & Sohn Verlag, Wiesbaden

Schroter S, Tite L, Smith R (2005) Perceptions of open access publishing: interviews with journal authors. Br Med J 330(7494):756, 4 pp (DOI: 10.1136/bmj.38359.695220.82)

Sinervo B, Lively C (1996) The rock-paper-scissors game and the evolution of alternative male. Nature 380(6571):240–243 (DOI: 10.1038/380240a0)

Smith JM (1972) Game theory and the evolution of fighting. In: Smith JM On Evolution. Edinburgh University Press, Edinburg, pp 8–28

Smith JM (1974) The theory of games and the evolution of animal conflicts. J Theor Biol 47(1):209–221 (DOI: 10.1016/0022-5193(74)90110-6)

Smith JM (1982) Evolution and the theory of games. Cambridge University Press, Cambridge

Smith JM, Price G (1973) The logic of animal conflict. Nature 246(5427):15–18 (DOI: 10.1038/246015a0)

Söllner F (2001) Die Geschichte des ökonomischen Denkens. Springer, Berlin

Sugden R (2002) Beyond sympathy and empathy: Adam Smith's concept of fellow-feeling. Econ Philos 18(1):63–87, available online at the URL: http://journals.cambridge.org/abstract_S0266267102001086

Szabó G, Fáth G (2007) Evolutionary games on graphs. Phys Rep 446(4–6):97–216 (DOI: 10.1016/j.physrep.2007.04.004), also available as arXiv preprint arXiv:cond-mat/0607344

Turner P, Chao L (1999) Prisoner's dilemma in an RNA virus. Nature 398(6726):441–443 (DOI: 10.1038/18913)

von Neumann J (1928) Zur Theorie der Gesellschaftsspiele. Mathematische Annalen 100(1):295–300 (DOI: 10.1007/BF01448847) also available online at the URL: http://resolver.sub.uni-goettingen.de/purl?GDZPPN002272717. Reprinted in: Taub AH (ed) (1963) Collected works of John von Neumann. Volume VI: Theory of games, astrophysics, hydrodynamics and meteorology. Pergamon Press, Oxford, pp 1–26. English translation: (1959) On the theory of games of strategy. In: Tucker AW, Luce RD (eds) Contributions to the theory of games: Volume IV. Annals of Mathematics Studies, vol 40. Princeton University Press, Princeton, NJ, pp 13–42

von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton, NJ

Weibull J (1995) Evolutionary game theory. MIT Press, Cambridge, MA

Zeeman EC (1980) Population dynamics from game theory. In: Nitecki Z, Robinson C (eds) Global theory of dynamical systems. Proceedings of an international conference held at Northwestern University, Evanston, Illinois, June 18–22, 1979. Lecture Notes in Mathematics, Vol 819. Springer, Berlin, pp 471–497 (DOI: 10.1007/BFb0087009)

Zermelo E (1913) Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. In: Hobson EW, Love AEH (eds) Proceedings of the fifth international congress of mathematicians (Cambridge, 22–28 August 1912). Cambridge University Press, Cambridge, vol II, pp 501–504, also available online at the URL: http://www.mathunion.org/ICM/ICM1912.2/. English translation: (2010) On an application of chess theory to the theory of the game of chess. In: Ebbinghaus HD, Fraser CG, Kanamori A (eds) Ernst Zermelo: Collected works/Gesammelte Werke: Volume 1/Band 1: Set Theory, Miscellanea/Mengenlehre, Varia. Schriften der Mathematisch-naturwissenschaftlichen Klasse der Heidelberger Akademie der Wissenschaften, vol 21. Springer, Heidelberg, pp 267–273 (DOI: 10.1007/978-3-540-79384-7)

AUTHOR QUERIES

AQ1. Please check whether the inserted citation for Table 5.1 is appropriate.
AQ2. Please check whether the inserted citation for Table 5.4 is appropriate.

Part III
Exemplary Model Applications

Chapter 6

Dynamic Scientific Co-Authorship Networks

Franc Mali, Luka Kronegger, Patrick Doreian, and Anuška Ferligoj

6.1 Introduction

Network studies of science greatly advance our understanding of both the knowledge-creation process and the flow of knowledge in society. As noted in the introductory chapter, science can be defined fruitfully as a social network of scientists together with the cognitive network of knowledge items (Börner et al. AQ1 2010). The cognitive structure of science consists of relationships between scientific ideas, and the social structure of science is mostly manifested as relationships between scientists. Here, we confine our attention to these relations. In particular, co-authorship networks among scientists are a particularly important part of the collaborative social structure of science. Modern science increasingly involves "collaborative research", and this is integral to the social structure of science. Ziman argues that the organizational units of modern science are groups and not individuals (Ziman 1994, p. 227).[1] Namely, co-authorship in science presents a

---

[1] Co-authorship in science is not the only form of scientific collaboration. de Haan (1997) suggests six operationalized indicators of collaboration between scientists: co-authorship; shared editorship of publications; shared supervision of PhD projects; writing research proposal together; participation in formal research programs; and shared organization of scientific conferences. As this list suggests, there are many cases of scientific collaborations that do not result in co-authored publications (Katz and Martin 1997; Melin and Persson 1996; Laudel 2002). Laudel (2002) reports that about half of scientific collaborations are invisible in formal communication channels either

F. Mali (✉) · L. Kronegger (✉) · A. Ferligoj
Faculty of Social Sciences, University of Ljubljana, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia
e-mail: franc.mali@fdv.uni-lj.si; luka.kronegger@fdv.uni-lj.si; Anuska.ferligoj@fdv.uni-lj.si

P. Doreianb
Department of Sociology, University of Pittsburgh, 2602 WWPH, Pittsburgh, PA 15260, USA

Faculty of Social Sciences, Kardeljeva pl. 5, 1000 Ljubljana, Slovenia
e-mail: pitpat@pitt.edu

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_6,

more substantial indicator than just scientific communication in one way or another. In continuation, we focus on the dynamics of different kinds of co-authorship networks.

Over the last 50 years, the study of the dynamics of co-authorship networks has been conditioned by the development of quantitative methodological approaches in various forms that include relatively simple descriptive statistics presented in time-series form, deterministic approaches, and stochastic agent-based modeling of network dynamics. We provide a brief overview of these approaches in this chapter. Many studies of co-authorship networks are typically described and understood in terms of very large networks involving tens of thousands of nodes. Science can be understood as social phenomena involving large numbers of scientists regularly performing specific actions that are consciously coordinated into large schemes (Ziman 2000, p. 4). Different disciplinary approaches allow the use of different statistical quantities to explain the topology of scientific networks. Some of the statistical quantities typically used to describe these networks are purely local. The other statistical quantities correspond to global descriptions. For example, the local property of a unit in the network is vertex degree, defined as the number of ties relating this unit to other units in the network. Corresponding global descriptions of the degree distribution, which is known to have a long tail for a wide range of different networks, can be constructed (see, for example, Lambiotte and Panzarasa 2009.).

Although co-authorship networks may provide a window on patterns of collaboration within science, they have received far less attention than have citation networks in bibliometrics (Newman 2004, p. 5200). There is a basic difference between co-authorship networks and citation networks. Citation networks are not personal social networks, even though they are, in part, the product of social network phenomena involving scientists. They do not capture the social interaction structure usually described in works on co-authorship networks. These social interaction structures are best described by co-authorship networks. The vertices of co-authorship networks represent authors, and two authors are connected by a tie if they co-authored one or more publications. These ties are necessarily symmetric. In citation networks, the vertices represent scientific productions,[2] and the links between them are directed citation ties from one scientific document to other such documents. In that sense, co-authorship networks contain much important information about cooperation patterns among authors as well as the status and locations of authors in the broader scientific community structures. The study of community structures through scientific co-authorship is particularly important

---

because they do not result in co-authored publications or in formal acknowledgments in scientific texts. In this chapter, we will use the term *collaboration* primarily to designate research that results in co-authored publications and other publicly available documents.

[2] We include papers, monographs, short articles, conference presentations, databases and patents within the term ‘scientific production.’

Table 6.1 List of major questions and models presented in this section

	Major issues addressed	Key answers/insights
Barabási and Albert (1999)	Ways of modeling cumulative advantage principle in co-authorship networks.	Using the preferential attachment model where a scale-free power-law distribution of the number of co-authors is a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to sites that are already well connected.
Watts and Strogatz (1998)	Ways of modeling the clustered structure of co-authorship networks at the macro level.	Small-world model overcomes the gap in clustering of real-world networks in comparison to random networks. Such constructed networks have small average shortest paths and incorporate clusterings (small dense parts of the network) which emerge in social networks.
Lorrain and White (1971), Doreian et al. (2005)	Ways of clustering the units in co-authorship networks regarding the structure of collaboration and representing the obtained clusters with their connections at the macro level.	The procedural goal of blockmodeling is to identify, in a given network, clusters (classes) of units (actors) that share structural characteristics defined in terms of some relation. Each such cluster forms a position. The units within a cluster have the same or similar connection patterns.
Snijders (1996), Snijders et al. (2010)	Ways of modeling the effects of actor characteristics and network positions on network evolution. Ways of modeling network dynamics and testing results using the inferential methods.	Stochastic actor-based modeling for network dynamics is based on longitudinally observed network data. It is meant to represent and model co-evolution of longitudinal network data and actor attributes, and evaluate the results within the framework of statistical inference.

because scientific (sub)disciplines might often display local properties that differ AQ2 greatly from the properties of the scientific network as a whole (Table 6.1).

This chapter is structured in the following way. Given that we treat co-authorship networks as social networks, we continue this introduction with a definition of a network. In the next section, we offer a brief historical overview of social network analysis with a focus on the dynamics of social networks. Section 6.3 contains an organizing typology of both the content and units of analysis for the topics we consider. Section 6.4 is the core of the chapter and provides an overview of known methodological approaches for studying dynamic scientific co-authorship networks. In the final section, we outline some benefits and limitations of each approach and finish with a statement of some open problems.

6.1.1 Networks as Graphs

A network consists of observed units and the relationships among them. Units can be represented as vertices and relationships (ties) as links. When using this skeleton representation, each network is a graph.

But this is a simplification of a network. Units (vertices) in the network can have properties. There can be multiple types of vertices in the network. An example is a social network where the vertices represent people and the groups to which they belong. Units also have many different properties (e.g. gender, age, income).

The links in networks can also be of different relational types and, further, strength of relationships can be indicated by adding weights. The vertices and links of networks studied in time have additional properties when time is considered. The timing of relational formations and dissolutions can be recorded and modeled. Duration of relational ties becomes another important property of relations when they are present.

The information of a graph can also be presented in a matrix form. The most common presentation is with the adjacency matrix in which there is a row and a column for each vertex. Non-zero entries in the matrix are present when links between two corresponding vertices exist.

Adjacency matrices can be extended further if we want to present more complex graphs. For example, if we want to present a graph with multiple links between the vertices, we associate the entry of a single cell a_{ij} in the adjacency matrix with the number of links between the vertices i and j. For the representation of valued graphs, which are graphs with values on the links, the value of a single cell a_{ij} in the adjacency matrix corresponds to the value on the link between vertices i and j.

6.2 A Brief History of Social Network Analysis

Histories of most entities usually have starting dates. However, establishing a starting date for an academic field is difficult because the contributing strands of ideas and methods for a field begin in different times and different places. Modern social network analysis (SNA) started when four distinct features were explicitly brought together (Freeman 2004). These features are: (i) a focus on structural matters by looking at actors embedded within a set of social relations and ties; (ii) the extensive use of systematic empirical data; (iii) heavy use of graphical imagery; and (iv) having foundations in formal, mathematical, and computational models. Recognizing the combination of these elements as defining social network analysis renders the establishment of a precise date of origin less than important. But, based on Freeman's narrative, a start date in the 1930s for what was to become SNA seems reasonable. What matters far more for the field are the operational ways in which the four core components are combined to help us understand network structures and processes.

Academic fields also require some social organization to support them in order to provide an accepted arena for the exchange of ideas and the development of an identity that nurtures a discipline. These were created for SNA within a span of 4 years. Barry Wellman established the International Network of Social Networks Analysts (INSNA) in 1976. He founded *Connections* a year later as a newsletter to distribute news, ideas, and information to members of the field. Lin Freeman established the flagship journal, *Social Networks*, in 1978. Finally, Russ Bernard and Alvin Wolfe started the annual Sunbelt Social Network Conference in 1980. All four entities have grown in size and influence since they were established. The European Network Conference was started in 1989, and in 1995 the two conferences were combined to form the Annual Sunbelt International Social Network Conference.

If we allow that SNA is what social network analysts do, it does not follow automatically that the field is coherent. Hummon and Carley (1993) examined all of the papers in the first 12 volumes of *Social Networks* to assess the state of the field and established that SNA was an integrated scientific community with a shared paradigm. They used ‘main-path analysis,’ a technique pioneered by Hummon and Doreian (1989, 1990) that helps study the citation patterns of a field. Hummon and Carley (1993) identified 6 main paths in the literature: (i) Role analysis and blockmodeling; (ii) Methods for network analysis; (iii) Concern with network data; (iv) Biased networks; (v) Attention to structure; and (vi) Analyses of personal networks. Of course, these paths for the movement of SNA intellectual ideas through the literature are linked. Hummon and Carley (1993) noted other features of the field. One was the heavy use of formal, mathematical, and quantitative methods. Another was the creation of substantive network ideas, and a third was the presence of prominent collaborative groups of social network analysts. All are consistent with the practice of ‘normal science’ in the sense of Kuhn (1996).

On looking at that list of main paths as intellectual foci for SNA, one feature leaps out by its absence: There is little about temporal issues[3] even though main path analysis is an explicitly temporal approach. Up until the beginning of the 1990s, SNA appeared to have had a profoundly static bias. The field’s concern was centered primarily – but not exclusively – on social structure and patterns of social structures. Given this, four event streams that can be dated as starting in the 1990s have changed the field dramatically.

The first was a series of three special issues of the *Journal of Mathematical Sociology* (JMS) that appeared in 1996, 2001, and 2003. All three issues, edited by Frans Stokman and Patrick Doreian, were devoted to “network evolution.” Based on the intuition that “network processes are series of events that create, sustain and dissolve social structures” (Doreian and Stokman 1997, p. 3), the three special issues had a series of papers that looked at network dynamics and network evolution using a variety of different formal models, simulation methods and statistical models.[4]

---

[3] This is consistent with the observations of Powell et al. (2005).

[4] Volume 30(1) of *Social Networks* (2010) was a special issue devoted to network dynamics that noted the importance of the three JMS special issues with papers building upon some of the earlier work.

The second event was the take-off of exponential random graph models (ergms) for the study of change in social networks. The origins of these models date from an earlier time, including the work of Holland and Leinhardt (1981) and Frank and Strauss (1986). One strand of this line work is founded on Wasserman and Pattison (1996) and Pattison and Wasserman (1999) and takes the form of p*-models. This forms the core of the software called Pnet (Wang et al. 2009), used for estimating ergms. Another strand features the work of Snijders (2001) and takes the form of SIENA (Snijders et al. 2010), also used for estimating ergms for studying the co-evolution of social actors and social networks. Yet another strand of related work is present in Statnet (Handcock et al. 2003) that includes the estimation of ergms. There has been a lively debate and an extensive cross-fertilization and collaboration between the groups centered at the University of Melbourne, the University of Groningen, Oxford University, and the University of Washington regarding ergms.

The third event is the movement of physicists into the realm of social networks, which also started in the 1990s. Bonacich (2004) labeled this as "the invasion of the physicists "in his review of Watts (2003) and Barabási (2002). To the extent that the physicists are inattentive to the substantive content of the SNA and reinvent old – and/or even square – wheels, this is an invasion. However, they also bring with them a variety of new modeling strategies and additional conceptualizations of network phenomena that include 'small-world' networks and 'preferential attachment,' two terms that have made fruitful entrances into SNA. The physicists have focused attention primarily on large networks with a view to delineating and understanding network topologies and dynamics.

The final event started in the early 1990s and resulted in the establishment of generalized blockmodeling (Doreian et al. 2005) as both a generalization and an extension of traditional blockmodeling, the main path in the SNA literature through 1992 identified by Hummon and Carley (1993). Thus far, this approach has been deterministic and not that attuned to network dynamics. Designed to delineate network structures through the use of an expanding collection of block types and types of blockmodels, it has the potential to contribute to the temporal delineation of fundamental network structures.

At face value, the four 'events' and the lines of active research that have followed them are different and could even be viewed as potential rivals. However, it will be unfortunate if they are seen in this fashion. Some of the ideas of physicists can be used to conceptualize mechanisms that can be incorporated into ergms to test these ideas with social network data. It is clear that the efforts of physicists to identify communities in networks have the same intent as blockmodeling. The work of Handcock et al. (2007) on discerning network structure through model-based clustering is also related, in intent, to blockmodeling, and it seems reasonable to couple, in some way, ergms and blockmodels. All of these four strands of research for understanding networks have been mobilized extensively since their first appearance. They have all emerged since Hummon and Carley's (1993) assessment and have the potential to be combined fruitfully in future research. While these streams of research are changing SNA to focus on network dynamics and network evolution, they do so while embodying all of the four defining features of SNA identified by Freeman (2004).

6.3 Levels of Analysis of Scientific Collaboration

6.3.1 Introduction

Understanding science as a social system implies considering science as fundamentally relational, and as a community-based social activity. "The collegian circles around a scientist refer to those local and distant peers or professional colleagues" (Schott 1993, p. 201). These collegian circles have several properties that vary from one scientist to another. Within social studies of science, there has been a strong interest in the spatial range of the collegian circle with attention given to local, national, or transnational scientific communities. These professional collegian circles in science have several other characteristics that are analytically distinct but, in reality, may be intertwined. Co-authorship networks in science have a "modular structure" (Lambiotte and Panzarasa 2009, p. 181). Understanding this modular structure of scientific networks is especially important because it helps account for the progress of science and the organization of scientific production within disciplinary frameworks. In reality, science never operates as a single community with hundreds of thousands of individual scientists. It is organized by many different networks that cut across the formal boundaries dividing science with regard to disciplinary, sectoral, and geographical levels. Of course, the membership of various networks overlaps considerably. These research networks are also in continuous processes of growth, decline, and dissolution (see, for example, Ziman 2000, p. 46 or Mulkay 1975, p. 519).

Classification of co-authorship networks can be done in several ways. Rogers et al. (2001) suggested a typology based on three features: (1) according to the units of the analysis, including individuals, teams of researchers, and R&D organizations; (2) according to the type of information used to develop the links between units – these might be based on interactions or information sharing or they could be based on positions of people in the social hierarchy; and (3) according to the institutionalized domains to which the authors belong, with an emphasis on intra-organizational or inter-organizational links between them. Sonnenwald (2007) suggested a more general classification to categorize various types of co-authorship networks: between researchers in university and industry sectors, between researchers in various scientific disciplines, and between researchers of various countries. In this section, we prefer to use another categorization, one adapting a suggestion by Andrade et al. (2009) who focused on three dimensions of co-authorship networks with their associated sub-dimensions of intra- and inter-dimensional co-authorship collaboration. The suggested dimensions are:

Table 6.2 Classification of levels of analysis of scientific collaboration

Dimension of the study		Examples of studies
Main dimension	Sub-dimension	
Cross-Disciplinary	Disciplinarity	Interaction links between Australian research networks (Rigby 2005), (see also: Wray 2002; Glänzel and Schubert 2004; Laband and Tollison 2000; Hornbostel 1997)
	Inter-disciplinarity	Interdisciplinary research analysis in French laboratories (Sigogneau et al. 2005) (see also Gibbons et al. 1994; Etzkowitz and Leydesdorff 2001; Qin et al. 1997; Braun and Schubert 2003)
Cross-Sectoral	Intramural	Academic research networks analysis (Lowrie and McKnight 2004; Wray 2002)
	Extramural	R&D cooperation models between industry and universities in Belgium (Veugelers and Cassiman 2005)
Cross-National	National	The interaction between immunology research institutes in Germany, due to their geographical location (Havemann et al. 2006)
	International	Comparative analysis of several countries of their international/national collaborated publications (Glänzel and Schubert 2005)

disciplinary with sub-dimensions of interdisciplinary and intradisciplinary, sector with intersector and intrasector, and geographic with international and intranational sub-dimensions. These are presented in Table 6.2.

6.3.2 *The Cross-Disciplinary Level*

For the cross-disciplinary level, given the presence of disciplinarity, there is a basic distinction between collaboration inside discipline (intra-disciplinarity) and collaboration between disciplines (inter-disciplinarity).

6.3.2.1 Disciplinarity

As stated in the introductory chapter of this book (see page xi et sqq.), "an academic discipline, or field of study, is a branch of knowledge which is taught and researched at the college or university level. Disciplines are defined (in part) and recognized by the academic journals in which research is published, and the learned societies and academic departments or faculties to which their practitioners [researchers] belong" (Börner et al. 2010). Many theorists of science have noted that all scientific disciplines are intellectually (cognitive) and socially structured (Fuchs 1992; Whitley 1984). Scientific disciplines represent institutional and organizational frameworks within which their intellectual products and cognitive styles are connected to the social structures, mode, and organization of the production of that knowledge. One of the basic characteristics of modern academic scientific communities is that they are still sharply differentiated and structured in terms of disciplines. Individual scientific disciplines can thus be seen as distinct intellectual and social organizational contexts.

Although co-authorship publishing is more common in the natural sciences than in the social sciences, it is continuously increasing in all main scientific areas (Wray 2002; Glänzel and Schubert 2004; Laband and Tollison 2000; Hornbostel 1997). Collaboration, operationalized through co-authorship, is now normative behavior and ubiquitous for practically all scientific disciplines (e.g., over 95% of articles in major periodicals in physics, biochemistry, biology and chemistry are co-authored (Braun-Munzinger 2009)).

6.3.2.2 Interdisciplinarity

In the last two decades, interdisciplinary collaboration has increased dramatically (see, for example, Gibbons et al. 1994; Etzkowitz and Leydesdorff 2001). This phenomenon is broadly discussed in Chap. 1 with attention focused on a tendency of modern science to form heterogeneous (interdisciplinary) teams of researchers solving pressing social problems and with higher accountability requirements (Börner et al. 2010). These attempts have been made to bridge narrow disciplinarities in science. An important feature stimulating interdisciplinary collaboration in modern science is the demand for innovations resulting from the juxtaposition of ideas, tools, and scholars from different scientific domains. Today, there is an overall agreement that inter-disciplinary links are vital for scientific progress because they have the potential to bring unprecedented intellectual and technical power. For example, the converging technologies of the NBIC fields (i.e. nanotechnology, biotechnology, information sciences, and cognitive sciences) are an example of new interdisciplinary research from fields that previously showed limited interdisciplinary connections (see, for example, Buter et al. 2010).

We know that different organizational and cognitive problems make the development of interdisciplinary research particularly challenging. Interdisciplinarity requires extensive networks of scientists and concepts, considerable time investments, and a need for researcher mobility between disciplines. As noted by Bordons and her collaborators, while collaboration among scientists from different disciplines is widespread, measuring it is not easy (Bordons et al. 2004, p. 441). Using bibliometrics, measurement of interdisciplinarity in publications can be approached from different perspectives that include co-authored publications among scientists from different disciplines, co-occurrence of several classification codes in publications, the interdisciplinary nature of journals, and the presence of cross-disciplinary references or citations. The most often used bibliometric indicator of such collaboration is the percentage of co-authored interdisciplinary publications. Yet, computing this percentage is affected by many factors, including the nature of the organization of scientific communities, R&D policy orientations, and the chosen operationalization of concepts (e.g., the classification scheme of disciplines that is used (Qin et al. 1997; Braun and Schubert 2003)).

6.3.3 The Cross-Sectoral Level

There is a basic difference between collaborations inside the academic scientific community (intramural cooperation) and collaborations between academic science, industry, and governmental bodies (extramural collaboration). Intramural networks in science are usually defined by collaboration within one department, research group, or institute. Extramural collaborations, on the other hand, consider also cooperation between different sectors (see, for example, Glänzel and Schubert 2004).

6.3.3.1 Intramural Collaborations (Intra-Sectoral Collaboration)

In modern science, the establishment of intra-mural networks is the result of the increased processes of professionalization of recent scientific activity. This has led to a large change in the organizational structure of science, and it's worth repeating Ziman's insight: "the organizational units of modern science are not individuals but groups" (Ziman 1994, p. 227). The organization of R&D activity in academic scientific institutions has created typical team structures – for example, modern research groups consist of principal investigators, co-principal investigators, junior researchers, post-docs, and doctoral students. Price suggested that research collaboration is, in part, a response to the shortage of scientists, which allows them to become "fractional" scientists (Price and Beaver 1966).

6.3.3.2 Extramural Collaborations (Cross-Sectoral Collaboration)

Cooperation between different sectors – academic science, industry and government – is now understood as the most important type of extra-mural collaboration. The concepts of 'Mode 2' and the 'Triple Helix' have extended the idea of research networking within and across sectoral borders. Both concepts were developed in the theory of science and R&D policy discussions after 1990. It seems that the concept of Mode 2 knowledge production presented in *The New Production of Knowledge* (Gibbons et al. 1994) became, in the mid-90s, the symbolic banner of new viewpoints regarding scientific collaborations across sectors. The authors of the new (Mode 2) production of knowledge linked the classical concept of transdisciplinarity, defined by common axioms that transcend the narrow scope of disciplinary worldviews through an overarching synthesis, with two additional factors: problem-driven research and research in applied contexts. Similarly, the concept of the Triple Helix has been developed as a neoinstitutional and neoevolutionary model for studying the networks across academic science, industry science, industry, and government sectors. In these networks, more important than the presence of the agents is the quality of their relationships in a given configuration (Etzkowitz and Leydesdorff 2001). Although there exists already an extensive expert literature on this type of cross-sectoral networks, there is still a lack of bibliometric studies dealing with co-authorship publications between academic and business-enterprise sectors (Lowrie and McKnight 2004, p. 436).

6.3.4 Cross-National Level

Networks of international collaboration have undergone dramatic structural changes in the last few decades. This is in contrast to intranational networks, where the intensity of collaborations have decreased (Hoekman et al. 2010; Glänzel and Schubert 2004; Katz 1994, see, for example).

6.3.4.1 National Collaborations

National collaboration, while visible in domestic contexts, is often regarded as less visible and treated as less important than international collaborations. Often, the observed (relative) high visibility and high citation attractiveness of internationally co-authored publications result in a kind of operational rule: international co-publications appear in high-impact journals and receive more citations than national papers (Glänzel and Schubert 2004). However, the overall visibility and international relevance sometimes does not necessarily reflect the impact of specific papers in solving specific problems at the local level. The results of national collaborations are often incorporated into publications dealing with trans-institutional and international co-authorship (e.g. Munshi and Pant 2004), and are focused directly on collaboration within a specific country (Gossart and Ozman 2009; Mali et al. 2010). Another important aspect of national collaboration results from the international orientation of bibliographic databases like the *Web of Science* or *Scopus*. Often, the results of national co-authorship and the resulting citation patterns, especially for smaller national scientific systems, are less visible in international bibliographic databases. This can be linked to inter-sectoral collaboration within nations. National collaborations across sectors have an additional complexity because they include the involvement of different administrative units. As a result, such research projects are complex and involve a wide range of different outputs of scientific production. Such complex information can only be reported qualitatively or measured through *local* information systems and electronic bibliographic systems; the Slovenian *COBISS*[5] and *SICRIS* databases[6] or the Turkish *ULAKBIM* database.[7]

---

[5] Co-operative Online Biographic System and Services, www.cobiss.si.

[6] Slovenian Current Research Information System, sicris.izum.si

[7] TURKISH ACADEMIC NETWORK and INFORMATION CENTER, www.ulakbim.gov.tr/eng/.

6.3.4.2 International Collaborations

In thinking about the spatial range of collaboration, there is an important difference between geographic distance and crossing international boundaries. While geographical distances between collaborative units in large nations can be long, the geographical distances between collaborating units in different countries can be short. Of the two, crossing international boundaries is more consequential than geographical distance with regard to scientific collaboration. While international scientific collaborations are important generally, they are especially important for small scientific communities such as, for example, the Slovenian scientific community. Isolated and parochial scientific communities are no longer a suitable environment for recognized scientific excellence. Indeed, it can be argued that they never were important in the history of science. Even in the early days of science, different forms of cooperation between scientists of different nations became important elements in the internationalization of science. Even so, because of the new forms of the globalized connections of science, "the traditional cosmopolitan individualism of science is rapidly being transformed in what might be described as transnational collectivism" (Ziman 1994, p. 218).

This trend of increasing international scientific collaboration through co-authorship is especially strong in recent decades. The number of internationally co-authored articles has risen at a faster rate than traditional 'nationally co-authored' articles (Wagner 2005). As noted in the expanding bibliometric literature, the level of international co-authorship is determined by many factors: the size of the country, 'proximity' between countries, either physical (geographical) proximity or immaterial proximity stemming from cultural affinity in a broad (historical, linguistic) sense, socioeconomic factors, changes in electronic forms of communication, and last but not least, the dynamics created by the self-interest of individual scientists pursuing their own careers.

6.4 Methodological Perspectives

6.4.1 Introduction

The development of methodological approaches for analyzing and modeling temporal scientific co-authorship networks has been founded on developments in graph theory and in SNA. To enable the discussion on temporal analysis of network properties, we describe some of the most relevant basic definitions of network properties that we need for understanding the content of coming sections (extensive explanations of SNA terminology and concepts can be found in Wasserman and Faust (1994)):

- *Degree* The degree of a vertex is defined as the number of ties linking this vertex to other vertices in the network. In lay terms, the degree represents the number of

co-authors for each researcher. As a global measure of the whole network, both the average degree or centralization can be considered.
- Network *density* is the proportion of ties in a network relative to the total number possible (sparse *vs.* dense networks).
- *Path* A path is a sequence of vertices and lines from initial vertex to the terminal vertex where all vertices different.
- *Path length* This is the number of ties it contains.
- A shortest path or a *geodesic distance* between two vertices u and v, denoted as l_{uv}, is the shortest path length between these two vertices. In co-authorship networks, the distance between two authors who collaborate is 1. As a global network characteristic, the average shortest path is usually considered.
- The *global clustering coefficient* can be viewed as the average probability of a tie between co-authors of a selected author. Technically, it measures the density of triangles in the network and therefore measures the extent of densely connected subgroups of vertices in the network.

Another important factor in the development of the field has been access to data sources on scientific collaboration. Before the development of electronic bibliographic databases and, especially, before the implementation of the scientific citation indexes initiated by Garfield (1955) this was a very difficult and time-consuming task. Some of the most visible electronic databases with academic content are the *Web of Science*, *SCOPUS* and *Google Scholar*. A broader discussion on databases and citation indexes can be found in Chap. 7 of this book.

The study of temporal networks, both with regard to network dynamics and network evolution, gained increasing attention since 1996. As noted in Sect. 2, special issues of the *Journal of Mathematical Sociology* (1996, 2001, 2003) were of value. We distinguish three basic approaches for studying dynamic scientific co-authorship networks: (i) basic analysis of network properties using temporal data (usually in the form of a time-series of snapshots, (ii) deterministic approaches to the analysis of scientific co-authorship networks, and (iii) statistical modeling of network dynamics.

6.4.2 Basic Analyses of Network Properties

One of the first analyses of temporal co-publication was presented by Zuckerman (1967) who studied the patterns of productivity, collaboration and co-authorship among Nobel Laureates. While her analysis was quite narrow, in the sense of focusing on a small elite among scientists, this was due to the limitations of the data available at the time. More than 20 years later, (Bayer and Smart 1991) focused on publication patterns of US PhD recipients in chemistry in 1960–1962. They used a longitudinal data set spanning from 1962 to 1985 to follow the careers of these researchers through time. Besides single-authored and multi-authored publications, they also distinguished dual-authorship and proposed a typology of publication patterns of scientists, including six categories which are highly correlated with co-authorship patterns. Researchers were categorized into groups of: Low producers, Burnouts, Singletons, Team Leaders, Team Players, Doubletons, and Rank-and-File types. With the development of electronic bibliographic databases, simple longitudinal analysis of network characteristics (including average vertex degrees, clustering coefficients, and density) became a common part of most studies of temporal co-authorship networks (see Babchuk et al. 1999; Glänzel et al. 1999; Kronegger et al. 2011a).

6.4.3 Deterministic Analysis of Dynamic Co-Authorship Networks

Although the time dimension is often included in the analysis of co-authorship networks, it has been mostly restricted to simple temporal time-series descriptions of some network characteristics and actor attributes. Such basic analyses can be found in a wide range of publications since results of practically every method for social network analysis can be represented in time as a series of snapshots. The most common goal of these methods is delineating structures within co-authorship networks and accounting for network properties by using some external parameters. Efforts of researchers to push the methodology further from simple description of differences between time snapshots are therefore rare and hard to find.

A fruitful way of delineating structures within co-authorship networks is to use blockmodeling procedures: Let U be a finite set of units and let the units be related by a binary relation $R \subseteq U \times U$ that determines a network $\mathbf{N} = (U, R)$. One of the main procedural goals of social network analysis is to identify, in a given network, clusters of units that share structural characteristics defined in terms of the relation R. The units within a cluster have the same or similar connection patterns to the units of other clusters. Result of clustering $\mathbf{C} = \{C_1, C_2, \ldots C_k\}$ is a *partition* of units $\mathcal{U}$ and relations R into *blocks* $R(C_i, C_j) = R \cap C_i \times C_j$. Each block is defined in terms of units belonging to clusters C_i and C_j and consists of all arcs from units in cluster C_i to units in cluster C_j. If $i = j$, the block $R(C_i, C_i)$ is called a *diagonal* block.

A *blockmodel* consists of structures obtained by shrinking all units from the same cluster of the clustering $\mathbf{C}$. For an exact definition of a blockmodel, we must be precise about which blocks produce an arc in the *reduced graph* and which do not. The reduced graph can be presented also by a relational matrix, called an *image matrix*.

The partition is constructed by using structural information contained in R only, and units in the same cluster are equivalent to each other in terms of R alone. These units share a common structural position within the network.

Blockmodeling, as a set of empirical procedures, is based on the idea that units in a network can be grouped according to the extent to which they are equivalent, in terms of some *meaningful* definition of equivalence. In general, different definitions of equivalence usually lead to distinct partitions.

Lorrain and White (1971) provided a definition of *structural equivalence*: Units are equivalent if they are connected to the rest of the network in *identical* ways. From this definition it follows that only four possible ideal blocks can appear (Batagelj et al. 1992b; Doreian et al. 2005)

$$\begin{array}{llll} \text{Type 0.} & b_{ij} = 0 & \text{Type 2.} & b_{ij} = 1 - \delta_{ij} \\ \text{Type 1.} & b_{ij} = \delta_{ij} & \text{Type 3.} & b_{ij} = 1 \end{array}$$

where δ_{ij} is the Kronecker delta function and $i, j \in C$. The blocks of types 0 and 1 are called the *null* blocks and the blocks of types 2 and 3 the *complete* blocks. For the nondiagonal blocks $R(C_u, C_v), u \neq v$, only blocks of type 0 and type 3 are admissible.

Attempts to generalize the structural equivalence date back at least to Sailer (1978) and have taken various forms. Integral to all formulations is the idea that units are equivalent if they link in equivalent ways to other units that are also equivalent. Regular equivalence, as defined by White and Reitz (1983), is one such generalization.

As was the case with structural equivalence, regular equivalence implies the existence of ideal blocks. The nature of these ideal blocks follows from the following theorem (Batagelj et al. 1992a): Let $\mathbf{C} = \{C_i\}$ be a partition corresponding to a regular equivalence $\approx$ on the network $\mathbf{N} = (U, R)$. Then each block $R(C_u, C_v)$ is either null or it has the property that there is at least one 1 in each of its rows and in each of its columns. Conversely, if for a given clustering $\mathbf{C}$, each block has this property, then the corresponding equivalence relation is a regular equivalence.

Until now, a definition of equivalence was assumed for the *entire* network and the network was analyzed in terms of the permitted ideal blocks. Doreian et al. (2005) generalized the idea of a blockmodel to one where the blocks can conform to more types beyond the three mentioned above, and one where there is no single a priori definition of ‘equivalence’ for the entire network.

The problem of establishing a partition of units in a network, in terms of a considered equivalence, is a special case of the clustering problem – such that the criterion function reflects the considered equivalence. Such criterion functions can be constructed to reflect the considered equivalence. They measure the fit of a clustering to an ideal one with perfect relations within each cluster and between clusters, according to the selected type of equivalence.

For the direct clustering approach, where an appropriate criterion function that captures the selected equivalence is constructed, a relocation approach can be used to solve the given blockmodeling problem (Doreian et al. 2005).

Inductive approaches for establishing blockmodels for a set of social relations defined over a set of units were discussed above. Some form of equivalence is specified, and clusterings are sought that are consistent with a specified equivalence. Another view of blockmodeling is *deductive* in the sense of starting with a

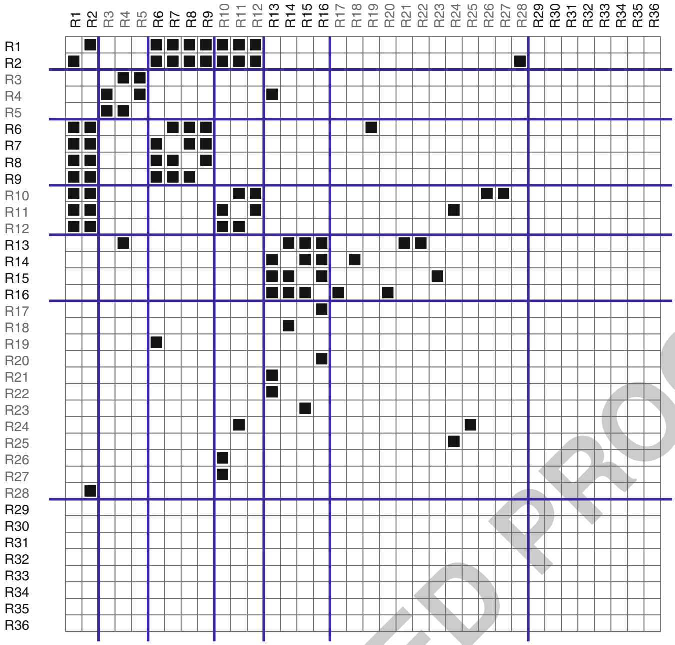


Fig. 6.1 An example of a blockmodel of a network with *multi–core–semi–periphery–periphery* structure.

this figure will be printed in b/w

blockmodel that is specified in terms of substance prior to an analysis. In this case, given a network, a set of types of ideal blocks, and a family of reduced models, a clustering can be determined which minimizes the criterion function. (For details, see, Batagelj et al. 1998; Doreian et al. 2005). Some prespecified blockmodels are designed as hierarchical models with the positions on paths linked by directed ties in a consistent direction. A core-periphery model is such a model where there is one (or several) core position that is strongly connected internally. Peripheral positions are all connected to core positions but not connected to each other, and they are not internally cohesive. There are variations of the core-periphery model; e.g., in which the periphery is not even connected to the core positions. All described blockmodeling approaches are implemented in the program Pajek (Batagelj and Mrvar 2010).

An example of the multi–core–semi–periphery–periphery structure is presented in Fig. 6.1. This specific structure, found in co-authorship networks, consists of: (i) simple cores comprised of scientists co-authoring with all, or most, colleagues in their core (units R3 to R5 and R13 to R16); (ii) bridging cores composed of researchers who connect two or more other simple cores (units R1 and R2); (iii) a semi-periphery made up of authors who co-author with proportionately fewer others in their position and have no systematic patterns of ties to scientists in other positions, and periphery of authors who do not co-author with other researchers from the network.

Several applications of blockmodeling of co-authorship networks have been published in recent years. For example, Said et al. (2008) distinguished several styles of co-authorship, including solo models (no co-authors), mentor models, entrepreneurial models, and team models. They conjectured that certain styles of co-authorship lead to the possibility of group-thinking, reduced creativity, and the possibility of less rigorous reviewing processes. Nooraie et al. (2008) examined co-authorship networks in three Iranian academic research centers in order to find an association between scientific productivity and impact indicators with network features. The collaboration networks within centers shared many structural features, including a "star-like" pattern of relations. Centers with more successful scientific profiles showed denser and more cooperative networks. Kronegger et al. (2011a) distinguished different co-authoring cultures in four scientific disciplines and delineated typical structures of scientific collaboration. They also extended blockmodeling by tracking locations, and hence positions, of authors across different time points.

Another effort to combine a static analysis of complexity at separate time moments with a dynamic analysis was presented by Erten et al. (2004) and by Gansner et al. (2005). They introduced a dynamic extension of multidimensional scaling (Richardson 1938; Torgerson 1952). Multidimensional scaling (MDS) is a set of data analysis techniques designed to display the structure of data in a geometrical picture. The algorithm of dynamic MDS is driven by the minimization of stress measured both within each analyzed year and over consecutive years by optimizing the resulting stress for a three dimensional array. This algorithm was recently implemented in Visone (Leydesdorff and Schank 2008) and used by Leydesdorff (2010) to study co-authorship networks, with additional information on co-word appearance and journal citation indexes. In this paper, he analysed the complete bibliography of Eugene Garfield for the years 1950–2010, graphically presenting its collaboration structure and citation dynamics around Garfields' work mainly dealing with the Science Citation Index.

6.4.4 Modeling Dynamic Scientific Co-Authorship Networks

Here, we present only an overview of modeling temporal co-authorship networks. Static models of macro-level network properties, which are based on stochastic rules of network generation, are discussed first. These have been mainly developed from graph theory by mathematicians and physicists who, with the development of the Internet in 1990, were interested in modeling accessible large real-world networks. The developments led from purely random graphs, built according to the Erdös and Rényi (1959) model, to small-world networks (Watts and Strogatz 1998), and to a range of models based on the concept of preferential attachment (Barabási et al. 2002; Newman 2000).

The idea of finding the rules fostering the growth and development of social networks, or as it was stated, modeling the real world graphs, was widely captured (mostly) by physicists. The basics for any kind of modeling of social networks were provided by the Erdös–Rényi random graph model, which is determined by a number of vertices (n) and the probability (p) that a link exists between two arbitrary vertices. Therefore, each random graph has approximately $p \cdot n(n-1)/2$ undirected links. A single vertex is linked to a binomially distributed number of neighbors. The limiting degree probabilities are Poisson distributed.[8]

The first generalization of the Erdös–Rényi random graph took the form of a *configuration model* where specific degrees are assigned (usually from a prespecified distribution) to all the vertices which are then randomly linked according to their degree. The construction of the model was proposed by Molloy and Reed (1995) and studied by many authors (see the overview provided in Newman 2003). This solved the problem of degree distribution in real-world graphs usually not having a Poisson distribution, as in the Erdös–Rényis random graph, but not the inability to model the clustered nature of empirical networks.

We consider also a very different approach to modeling social network dynamics, one which returned to and is founded upon ideas within social science. The approach of the physicists has been intent on reproducing the topological form of real-world networks, and it proposes some generic processes of growth and change while ignoring an extensive tradition of sociological and psychological knowledge regarding the behavior of individuals. This alternative (more sociological) approach focuses on single actors and their involvement in the smallest possible social unit of analysis, the dyad. This type of modeling is labeled 'stochastic actor-based modeling' (Snijders 1996). Its purpose is to represent network dynamics on the basis of observed longitudinal data in the form of explicit models and to evaluate them (or a family of models) within the paradigm of statistical inference. This implies that the models are able to represent network change as the result of dynamics being driven by many different tendencies, especially structurally based micro-mechanisms. These mechanisms can be theoretically derived and/or based on empirically established properties in earlier research. Of great importance is that these mechanisms may well operate simultaneously (Snijders et al. 2010). One limitation of these models is that they are restricted to a smaller predetermined number of actors and do not directly consider more global mechanisms of network growth.

6.4.4.1 Modeling "Real-World" Networks

Social studies of science have long had an interest in linking scientific production to the network structures of scientific communities. Different models have been proposed as representations of processes driving co-authorship (as collaboration) in

---

[8] Mathematical notations of models in this section are based on those used by Kejžar (2007).

science that help account for the form of large-scale scientific networks and predict scientific production. One contains an argument that if scientists from particular scientific disciplines (specialties) collaborate with others inside their disciplines, then we would expect to find distinct clusters in the knowledge-production network – exactly the clustering noted in many empirical networks – and this would correspond to small-world network structure (as described below). Alternatively, if the network was generated by preferential attachment (see below) as a mechanism – *where young scientists publish with well-established scientific stars* – then we would expect to find a scale-free network structure whose degree distribution satisfies a power-law. If the network is based on a cross-topic collaboration, then we would not expect to find strong fissures in the network, but instead find a structurally cohesive network (Moody 2004). All of the above-mentioned network structural processes lead to specific dynamics for scientific networks that, in turn, generate distinctive network structures or topologies. These models for generating the structures of large-scale and complex networks can be expected to hold also for co-authorship networks in science. Large-scale co-authorship networks can have local (such as clustering) structural properties as well as global (such as average distance between nodes) structural features. Local and global characteristics of networks help to define network topologies such as "scale-free networks" and "small-world networks." These network topologies are the result of network-generating processes and can lead to further dynamics of these networks in different ways. For example, the principle of preferential attachment to vertices of higher degree leads to a dynamic where "the-rich-get-richer. "In the case of science, this implies that those scientists who experience early success gain higher shares of subsequent rewards. We next consider scale-free and small-world science network structures in more detail.

6.4.4.2 The Small-World Model

The small-world network structure of scientific co-authorship implies network forms where the level of local clustering (one's collaborators are also collaborators with each other) is high, but the average number of steps between clusters is small. In these small-world networks, internal ties to clusters tend to form more cohesive clusters within boundaries, as compared to the more extensive and less cohesive overall networks that include their external ties. According to various social network analysts, the small-world model was inspired by the work of de Sola Pool and Kochen (1978) who partially formalized the much more famous application of Travers and Milgram (1969). It expresses the simple idea that any two individuals, selected randomly from almost anywhere on the planet, are 'connected' via a path of no more than a small number of intermediate acquaintances. The (limited) empirical evidence suggested that this small number is about 6. This notion became a popular idea in the Broadway play named *Six Degrees of Separation*. The first practical evidence for the existence of a small-world phenomenon was first provided by the psychologist Milgram (Berg 2005, p. 46). Milgram's experimental result was

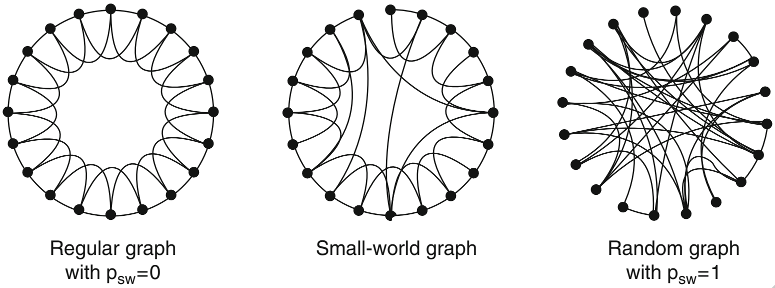


Fig. 6.2 Small-world structure simulation with different levels of randomness

regarded as a good starting point for analyzing the underlying structure of scientific co-authorship.

Later, Watts and Strogatz (1998) formally defined the small-world model in order to construct networks with the following properties that mirror some observed social networks: (i) having short paths between any two vertices (and hence, smaller average lengths for the shortest paths) and (ii) also incorporates clustering (small dense parts of the network). Knowing that geographical proximity of vertices plays a role in the formation of links (especially for humans), they considered a ring-lattice with n vertices. Each vertex had m_{sw} edges to its neighbors. Then they rewired each edge with a probability p_{sw} by relinking the second end of the edge to a randomly chosen vertex. The probability p_{sw} enables this network to vary from an ordered, finite dimensional lattice to a completely disordered network. The ring-lattice does not show a small-world effect since the average shortest path grows faster than a logarithmic rate of increase with the number of vertices, but it has strong local clustering. When the edges are rewired, Watts and Strogatz noticed that replacing a few long-distance connections hugely reduced the network's average shortest path and, as a result, a small-world effect appears. When $p_{sw} = 1$, the network becomes completely disordered where local clustering is no longer present and the average shortest path is small. Watts and Strogatz showed, by numerical simulation, that there is a relatively large p_{sw} interval in between the two extremes, for which the model exhibits both low path lengths and clustering (Fig. 6.2).

AQ3

Newman (2001, 2004) provides an excellent overview of the analysis on the topology of small-world network structures, highlighting key organizing principles that guide ties among the nodes in the network. According to Moody (2004), an archetypal small-world network will have many distinct clusters, connected to each other by a small number of ties. An analysis dealing with the dynamics of co-authorship publication networks in Slovenian sociology (Mali et al. 2010) showed that, to some extent, they conform to the small-world network structure: there are groups of sociologists that are very connected inside small groups but connected with others in non-systematic ways. Further results, obtained by using the blockmodeling approach, pointed to a publication strategy of those sociologists in Slovenia who are included in these small-world structures and are more oriented to parochial scientific reports or publications in Slovene. Consistent with this, they publish less in the international peer-reviewed journals than the sociologists outside this small-world structure. The results of these empirical analyses of Slovenian sociologists suggest that the presence of a too ‘closed’ and dense co-authorship network in science can have negative effects on the international orientations of scientists in a small scientific community. This implies that, for scientific performance and scientific excellence, it is much more important to have ‘open’ networks that have many structural holes (gaps between actors that create opportunities for brokerage). This is especially important for linking micro-level interactions (cooperation inside internal scientific organizations) to macro-level patterns (cooperation in the international scientific community). Burt provided evidence suggesting that new ideas in society emerge from selection and synthesis processes that operate across structural holes between groups. Positive performance evaluations and good ideas are disproportionately in the hands of people whose networks span structural holes. The ‘between-group brokers’ are more likely to have ideas viewed as valuable (Burt 2004) within the community.

6.4.4.3 The Preferential Attachment Model

The scale-free network structure, in one version or another, corresponds fairly closely to the sociological model of cumulative advantage in science. The first systematic representation of this model was provided by Merton (1973). Following Merton, there was a research stream in the literature that invoked the idea of cumulative advantage as a central explanatory principle for the social stratification of science. Merton’s studies were concerned with both organizational and functional aspects of science as an institution capable of self-regulation. This approach found its most significant (or at least most famous) expression in the description of the normative structure of science. Merton focused his attention on four institutional imperatives: universalism, communism, disinterestedness, and organized skepticism. Merton and other scholars working within institutional approaches (including Barber, Zuckerman, and Hagstrom) analyzed how norms regulate scientific activity. They studied the ways in which resources and rewards (including scientific prestige and opportunities to publish) are assigned and distributed within the scientific community (see, for example, Matthew 2005; Bucchi 2004).

The idea of cumulative advantages comes from the passage in Matthew’s Gospel: “For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.” (Hence the term “the Matthew effect.”) Translating the idea of cumulative advantage in science implies that those scientists who already occupy a position of excellence are rewarded far more than others in their field. Scientists who are rich in recognition find it easier to obtain additional recognition. In contrast, scientists who receive little recognition for their research efforts have reduced chances for future recognition.

Merton argued that cumulative advantage is a primary mechanism in modern science for the creation of scientific stars.[9]

A more quantitative and bibliometric basis for assessing the phenomenon of unequal distribution of publications (in connection with the unequal distribution of awards) in modern science has been provided also by Price (1976; 1963) in the form of his measure of scientific productivity. According to Price's law of scientific productivity, "...half of the scientific papers published in a given sector are signed by the square root of the total number of scientific authors in that field" (Price 1963, p. 67). This means that a relatively small number of highly productive researchers are responsible for most scientific publications. Price's law is founded on the same probabilistic basis as the earlier established Lotka Law,[10] the Bradford Law,[11] and Pareto and Zipf[12] distributions.

Both Price's law and the Matthew effect depict the scientific community as a structure characterized by marked inequality and a heavily pyramidal distribution of scientific rewards and publications. They are linked by the principle of preferential attachment which contains, for the case of scientific co-authorship networks, two generic aspects: (1) the continuous addition of new vertices into the network system and (2) preferential connectivity of new vertices. It means that a common feature of the models of scientific co-authorship networks, based on the rationale of preferential attachment, continuously expands by the addition of new vertices that are connected to the vertices already present in the networks. Additionally, in these models a new actor is, at best, most likely to be cast in a supporting role with more established and better-known actors. Further, no scientific field expands with

---

[9] Merton and his sociological followers (see Allison et al. 1982; Cole and Cole 1973) have analyzed several other similar mechanisms with regard to science networks, collaboration structures, and recognition in science:

1. The "halo effect" in science denotes the advantage of scientists in more favorable institutional locations.
2. The "Matilda effect" points to the discrimination against the participation of women in scientific activity.
3. The "gatekeeper" labels those scientists who can influence the distribution of resources such as research funds, teaching positions, or publishing opportunities because they occupy key decision-making positions within scientific institutions.
4. The idea of an "invisible college" was introduced on the basis of a seventeenth century expression denoting informal communities of researchers that cluster around specific projects or a research theme and that often turn out to be more influential in terms of knowledge production than formal communities (departments, research centers, scientific committees).

[10] Lotka's law states: The number of authors making n contributions is about $1/na$ of those making one contribution, where a is often about 2.

[11] Bradford's law states: Journals in a field can be divided into three parts: (1) a core of a few journals, (2) a second zone, with more journals, and (3) a third zone, with the bulk of journals. The number of journals in these three parts is $1 : n : n^2$.

[12] Zipf's law states: The probability of occurrence of words or other items starts high and tapers off. Thus, a few occur very often while many others occur rarely. The formal definition is: $P_n \sim 1/n^a$, where P_n is the frequency of occurrence of the nth ranked item and a is close to 1.

an endless growth of new vertices but is constrained by the operation of feedback effects.[13] It follows that there exist nodes, called "hubs" or "Angelpunkten oder Naben" (Berg 2005, p. 53), that acquire more links than another nodes. In such types of networks, preferential attachment and the system feedback dynamics play very important roles.

Crane (1972) provided an analysis of (global) scientific networks where informal members of scientific elites (in Moody's terminology, scientific stars) through whom the communication of scientific information both within scientific disciplines and across scientific disciplines is directed have the position of "hubs". Namely, they are central scientists in the network from where the information is transferred to all other scientists in the network. They also communicate intensively with each other. The idea of scientific networks with hubs can be used as a starting point to relate micro-level interactions (for example, in a local/national scientific community) to macro-level patterns (for example, the global scientific community). Through the informal groups of scientific elites, the small-scale interactions become translated into large-scale patterns. These large-scale patterns (international science) also have feedback effects on small groups (parochial/national science). The production and diffusion of the most creative and excellent scientific ideas in the world arise from the brokered networks (Granovetter 1973, p. 1360).

Albert and Barabási (2001) provide examples of many real-world networks whose degree distributions are far from a Poisson distribution. They showed that distributions can be approximated with a power-law function. They proposed a new evolving network model – PA or preferential attachment model (Barabási and Albert 1999). The model was presented as one that "shifts from modeling network topology to modeling the network assembly and evolution" (Albert and Barabási 2001). The idea behind the model was to capture the construction (development) of a network that could possibly explain the large number of observed power-law degree distributions in real networks. Before, there existed mostly network models with a fixed number of vertices among which links were added according to a particular procedure (process). Since real networks typically grow with the addition of new links and vertices that are not added randomly, Albert and Barabasi included the following ideas in their model.

The algorithmic statement of their model, given a set of vertices in a network, consists of the following two processes in a sequence of steps:

- At every time step, a new vertex v is added to the network.
- m_{ba} edges are created from the new vertex v to the vertices that are already in the network. These vertices are chosen with a probability proportional to their current

---

[13] (Berg 2005, p. 54) points out that "the effect of the positive feed-backs, namely, the advantages of old nodes against new ones as well as the attractiveness of the already networked nodes for newly added ones are leading to the growth of networks based on the preferential attachment", ("...doch in einem bestimmten Bereich sind positiven Rueckkopplungen feststellbar. Beide Effekte zusammen, der Vorteil, den alte Knoten gegenueber neuen haben sowie die Attraktivitaet besonders vernetzter Knoten fuer neu hinzukommende, fuehren dazu, dass das Wachstum des Netzes einer bevorzugehenden Verbindungswahl folgt.")

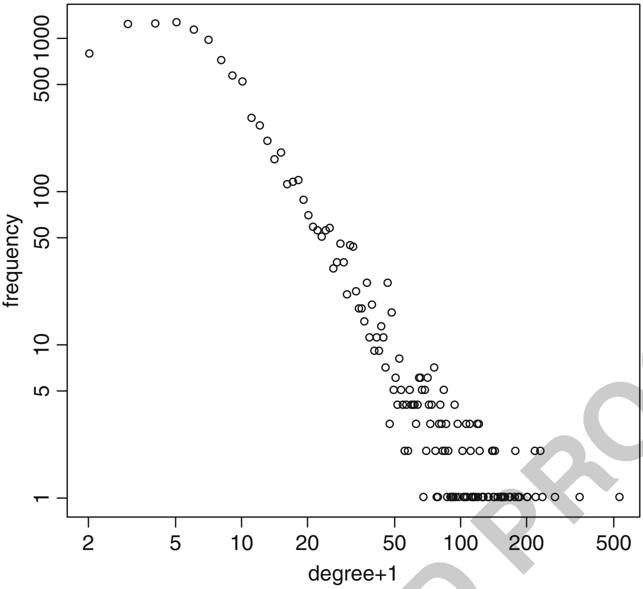


Fig. 6.3 Degree distribution in a co-authorship network of Slovenian researchers (Kronegger et al. 2011a) presented on a log-log scale. A large number of researchers with a small number of co-authors and a small number of researchers with a high number of co-authors indicates the existence of a preferential attachment mechanism in the process of network growth

degree. The probability of choosing vertex u can be written by $k_u / \sum_j k_j$ (where k_u represents the current degree of vertex u).

After t time steps, there are $t + m_0$ vertices in the network (where m_0 denotes the number of vertices at the beginning of the process) and tm_{ba} edges. It was first shown with simulations that the degree distribution of the whole network resulting from the operation of this model follows a power-law distribution with an exponent $\gamma = 3$ (Fig. 6.3).

Such scale-free networks as these generated through the principle of preferential attachment, in addition to not having a Poisson distribution of links around nodes, also have the interesting property of being very resistant to random attack. Almost 80% of the links can be cut before a scale-free network is destroyed, while the corresponding percentage for an exponential network is less than 20%.

Many generalizations about preferential attachment models have been made (Albert and Barabási 2001; Newman 2003). Systematic divergence from the power-law distribution at small degrees can be seen in many real-world networks. Therefore, Pennock et al. (2002) proposed incorporating a mixture (weighted addition) of preferential attachment and random attachment in the model. A further refinement of this model, where a directed version of the model was taken into account, is implemented in Pajek (Batagelj and Mrvar 2010). There, at each step of the growth a new vertex is selected according to its weighted in-/outdegree and some uniform attachment.

Another generalization about both small-world and preferential attachment, developed for two-mode networks, comes from Latapy et al. (2008) who present a nice overview of method developments for two-mode networks. Opsahl (2010) provides another attempt to overcome the issues of higher clustering coefficients in projections of two-mode to one-mode networks by redefining both the global and local clustering coefficients so that they can be calculated directly for two-mode structures.

6.4.4.4 Applications Featuring Co-Authorship Networks

Newman (2001) showed that collaboration networks form small-worlds in which randomly chosen pairs of scientists are typically separated by only a short path of intermediate acquaintances. He further provided information on the distribution of the number of collaborators, demonstrated the presence of clustering in the networks, and highlighted the number of apparent differences in the patterns of collaboration between fields. Also, Newman (2004) used data from three bibliographic databases for biology, physics, and mathematics to construct networks in which the nodes were scientists. He used these networks to answer a broad variety of questions about collaboration patterns, how many papers did authors write and with how many people, what is the typical distance between scientists through the network, and how do patterns of collaboration vary between subjects and over time.

Barabási et al. (2002) analyzed co-authorship data from electronic databases containing all relevant journals in mathematics and neuroscience for the period between 1991 and 1998. They found that network evolution is governed by preferential attachment. However, contrary to their predictions, the average degree in the networks they analyzed increased, and the node separation decreased in time. They also proposed a model that captured the network's time evolution.

Moody (2004) made an important contribution by identifying several types of individual scientific collaboration behavior that leads to the development of co-authorship networks that resemble networks generated according to the principles of small-world and preferential attachment. Recently, several articles that test the principles of small-world and preferential attachment have been published. Some are based on local databases like the Slovenian *COBISS* (Mali et al. 2010), while others use general databases like *Web of Science* (Perc 2010; Wagner and Leydesdorff 2005; Tomassini and Luthi 2007).

6.4.4.5 Developments of Models for Longitudinal Network Data

After the pioneering work of Erdös and Rényi on random graphs, and after the first applications of graph theory appeared in the sociological community (de Sola Pool and Kochen 1978), one group within the scientific community moved *away* from the idea of merely reproducing some global properties of "real-world" network properties. Instead, they focused on an approach designed to include micro-mechanisms that generate local changes in networks that also help account for the macro-structure of networks. Moreover, these efforts were designed to treat the micro-mechanisms as hypotheses that could be evaluated through statistical inference. The basics for these models of network change are (as already mentioned in the previous section) random graphs and random graph processes which incorporate the probabilistic uncertainty into the models. Uncertainty is present because there are many potential generators for observed graph structures, including co-authorship. From a methodological perspective, modeling the dynamics in social networks led to several obstacles. Probably the most persistent one was the interdependencies of the units comprising the networks. This problem remained untouched for almost 20 years. Indepth overview of approaches and methods to modeling network changes in time can be found in Frank (1991), Snijders (1996), and Snijders et al. (2010).

There are two distinct approaches to modeling network changes in time: models that implement change in discrete time steps, and more advanced models where time is modeled by continuous flows. Success in modeling change in social networks began in 1959 when Katz and Proctor showed that change in preferences for making ties in the network could be represented by a stationary, discrete-time Markov model. Of course, they assumed the independence of dyads within which all the modeling took place. In 1981, Holland and Leinhardt published a very influential article on log-linear models of network change which initiated a vigorous research stream devoted to the development of a broad class of models. One basic model, called p_1, was developed by Fienberg and Wasserman (1981) and Wasserman and Weaver (1985). Authors also provided efficient algorithms to find the maximum-likelihood estimators of parameters defining appropriate probability functions. Fienberg et al. (1985) showed how to handle social network data with the Holland-Leinhardt model and its extensions in contingency tables by using basic log-linear models. The longitudinal dimension to the log-linear approach was added by Wasserman (1987) and Wasserman and Iacobucci (1988).

Conditionally uniform models (Holland and Leinhardt 1975) are often used for modeling directed graphs where the probability distribution for forming new ties is uniform, conditional upon a certain set of attributes. In these models, the conditional statistics are defined by attribute variables and contain the most relevant effects of the studied phenomena, while the rest is explained by random factors. Conditionally uniform models become very complicated when more informative conditioning on attribute variables is included into the model. Such models for longitudinal binary network data at 2 time points – conditional upon the entire network at the first time point, and upon the numbers of newly formed and dissolved ties for each actor – were developed by Snijders (1990). The idea of conditioning the changes in the network on the first measured network resolves most of the unexplained factors that determined the development of network before its first measurement.

Modeling changes in continuous time with Markov chains was adapted by Coleman (1964) to tackle some classical sociological problems. Holland and Leinhardt (1977) extended this idea to model networks of interpersonal affect between actors. They developed a valued Markov chain approach to model the process by which social structure based on affect influenced individual behavior.

The basic assumptions underlying the use of the continuous time Markov chain model are:

1. Between the observation moments, time runs continuously. Changes can be made (but are likely to be unobserved) at any moment, t.
2. The network $X(t)$ is the outcome of a Markov process.
3. At each single moment, only one relational tie or variable attribute may change.

Wasserman (1978, 1980a,b) continued this approach and provided estimators for parameters of various models. He started with a simple model of reciprocity in directed graphs, but without complicated dependencies between ties such as those generated by transitive closure.

The breakthrough in modeling the dynamics in social networks was the relaxation of the assumption of conditional independence between dyads (Mayer 1984). This was an important step since most sociological theories assume at least some kind of dependence structure between dyads. Another important step came in the form of dropping the stationarity assumption (Leenders 1995). Leenders also developed a mechanism to allow changing rates for all dyads to be dependent on arbitrary covariates, with the assumption that these remain constant between the observations.

In recent years, these models became known as stochastic actor-oriented models which have been developed to consider a variety of micro-mechanisms for generating network structure. These models are based on an assumption that each actor has his/her own goals which he/she tries to advance in accordance with his/her constraints and possibilities. Snijders (1995) referred to this approach as ‘methodological individualism’ where the driving force behind the network dynamics comes in the form of actions by actors.

Each attempt to model specific sociological problems or theories produced a new mathematical model that filled the gaps along the way to obtaining a better representation of reality. Yet an important feature still had to be addressed because most of these models lacked an explicit estimation theory.

The first models addressed some basic questions. A baseline of development can be followed through the work of several authors. Jackson and Wolinsky (1996) presented a model where the benefits and costs of ties affected the evolutionary trajectories of networks and the form of equilibrium structures. Hummon (2000) constructed actor-oriented simulation models of ‘Jackson and Wolinsky actors’ to study temporal network dynamics. He specified choices under four combinations of tie formation and deletion rules: unilateral and mutual tie formation, and unilateral and mutual tie deletion. This process generated eight types of networks: Null, near-Null, Star, near-Star, Shared, near-Shared, Complete and near-Complete as equilibrium structures. Doreian (2006) provided a formal proof via exhaustive examinations of the structures identified by Hummon (but only for tiny networks), and this line of work was extended by Xie and Cui (2008a,b). In another line of development, Marsili et al. (2004) presented a simple model using the creation of links to friends of friends, a mechanism that was introduced by Vázquez (2003) in the context of growing networks. This model is similar to the one proposed by

Davidsen et al. (2002) which explained the emergence of the small-world property in some social networks.

In the model of Skyrms and Pemantle (2000), individual agents begin to interact at random, with the interactions modeled as games. The game payoffs determine which interactions are reinforced, and network structures emerge as a consequence of the dynamics of the agents' learning behavior.

More complex network dynamic models with larger but still quite restricted numbers of tendencies were presented by Jin et al. (2001). They propose some simple models for the growth of social networks based on three general principles: (i) meetings take place between pairs of individuals at a rate that is high if a pair has one or more mutual friends and low otherwise; (ii) acquaintances between pairs of individuals who rarely meet decay over time; (iii) there is an upper limit on the number of friendships an individual can maintain. Their models incorporate all of these principles and reproduce many of the features of real social networks, including high levels of clustering or network transitivity and strong community structure in which individuals have more links to others within their community than they have to individuals from other communities. The important feature of their models is the inclusion of a time scale on which people make and break social connections.

6.4.4.6 Simulation Investigation for Empirical Network Analysis – Siena

The problem of inference in modeling dynamics of social networks on the basis of the observed longitudinal data was addressed by Snijders (1996) and extended further by Snijders et al. (2010). These models are based on longitudinal data and include representations of network dynamics as being driven by many different tendencies. These include micro-mechanisms, which have been theoretically derived and/or empirically established in earlier research, and which may well operate simultaneously. One of the most important characteristics of these models is the evaluation of their results within the paradigm of statistical inference, which makes them suitable for testing hypotheses and estimating tendencies that drive tie formation and dissolution at the level of individual units using reciprocity, transitivity, homophily, etc.

The model assumptions are:

- The model is basically defined for directed relations. In the case of undirected networks (e.g., co-authorship networks) the tie formation is additionally modeled using different mechanisms (e.g., a unilateral forcing model, unilateral initiative, and reciprocal confirmation, etc.)
- The network is observed in 2 or more discrete timepoints. But the underlying time parameter in the model is continuous.
- Changes in the network are outcomes of a Markov process, which means that the change in the network from one state in time point t_i to new state in time point

t_{i+1} is conditioned only to the state of the network in time point t_i. The process does not take into account any other historical events.

- The actors control their ties, which means that changes in ties are made by actors who send the tie on the basis of their and others' attributes, their position in the network, and their perceptions about the rest of the network. Regarding the last, it is assumed that actors have full information about the network and the other actors.
- At any given moment, only one probabilistically selected actor may get the opportunity to change only one tie.

The actor-based process is decomposed into two stochastic sub-processes:

1. The change-opportunity process models the frequency of the tie changes by actors. The opportunity to change the tie depends on the network locations of the actor (e.g., his or her centrality) and on actor covariates (e.g., gender or age).
2. The change-determination process models the change of the tie when an actor gets an opportunity to make a change. The change of the tie can be made with equal probabilities or with probabilities depending on attributes or network positions. Perceived attributes and position (the environment) of the actor is included into the actor's objective function, which expresses how likely it is for the actor to change his or her network environment in specific way (i.e., initiate, withdraw tie, or keep the present situation).

To use this model with observed data means that parameters have to be estimated by some statistical procedure. Since the model is too complicated for classical estimation methods such as maximum likelihood, Snijders (1996, 2001) proposed a procedure using the method of moments implemented by a computer simulation of the network change process. The procedure he proposed uses the first observation of the network as the (unmodeled) starting point of the simulations. This implies the estimation procedure is conditioned on the first observed network of a series of observations of that network.

The limitation of such models is that they are limited to a predetermined and rather small number of actors (between 100 and 200 actors) and do not directly consider the mechanisms of network growth. The methods and algorithms developed by Snijders et al. (2008) are implemented in the computer package SIENA.

Stochastic actor-based modeling of network dynamics was initially developed for modeling the change in directed networks. The undirected networks such as co-authorship networks are a special case where reciprocity cannot be used as a mechanism of network change. Although several articles have been published using SIENA models, to our knowledge, only Kronegger et al. (2011b) dealt with undirected networks to study the dynamics of co-authorship networks of Slovenian researchers working in physics, mathematics, biotechnology, and sociology in the time period from 1991 to 2005. In their study, they operationalized the modeling of global network parameters used in the preferential attachment and the small-world models with stochastic actor-oriented modeling.

6.5 Summary

Access to bibliographic databases and the availability of powerful quantitative social network approaches increased the number of studies of co-authorship networks in different scientific fields. There are several classification schemes for analytical approaches to analyzing the dynamics of co-authorship networks. We decided to classify them according to the types of models. The first type of model provides the basic analysis of whole co-authorship network properties. Such network characteristics are degrees, clustering coefficients, and density. The usual statistical approach used in these models is time-series analysis of listed properties.

Deterministic models (the second type) and stochastic models (the third type) are usually used to analyze actor-based co-authorship networks and attribute characteristics. To study the structure within the co-authorship networks, blockmodeling approaches are recommended. To model dynamic co-authorship networks, several approaches can be used according to the chosen level of analysis. Models on the macro level (whole network level) were mostly developed by mathematicians and physicists. These are models of "real-world" networks, small-world models, and preferential attachment models. The alternative stochastic actor-based model (implemented in SIENA) was developed by social scientists and statisticians. This model focuses on single units and on dyads. This powerful model studies network change in time as the result of micro-mechanisms for generating the network structure.

There are several indicators that show a huge development of analytical approaches to studying social networks through time. The powerful stochastic actor-based networking model has one disadvantage in that it can only be used to analyze a few hundred units in the network. Therefore, there is a need for similar models to analyze large networks.

Key points

Modeling of co-authorship networks can be approached in terms of the different perspectives and goals that have been outlined in this chapter. As a partial summary, the following items are important:

1. Level of the analysis: the macro level (whole network) or the micro level (unit). Which one is used depends on the goal(s) of the study. There are the following three variants:
    a. Describing the topology of the macro structure
    b. Understanding the micro-level changes at the actor level
    c. Coupling the micro-level processes to the generation of the network's macro structure.
2. Size of the network: some models can process only a limited number of units (e.g. stochastic actor-based modeling and direct blockmodeling),

while others can handle large networks (e.g., preferential attachment, the small-world model, and indirect blockmodeling).

3. Discrete-time models (e.g., blockmodeling) or continuous-time models (e.g., stochastic actor based modeling).
4. The analysis of the evolution of co-authorship networks only (e.g., small-world model, preferential attachment, blockmodel) or including external characteristic of network (e.g., scientific field) and/or actor attributes (e.g., age or gender of researcher) using modeling approaches (e.g., stochastic actor based modeling).
5. Needs of graphical representation of co-authorship network evolution (e.g., preferential attachment, blockmodeling, multidimensional scaling).

References

Albert R, Barabási AL (2001) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97 (DOI: 10.1103/RevModPhys.74.47), also available as arXiv preprint arXiv:cond-mat/0106096

Allison PD, Long JS, Krauze TK (1982) Cumulative advantage and inequality in science. Am Sociol Rev 47(5):615–625 (DOI: 10.1146/annurev.soc.32.061604.123127, stable JSTOR URL: http://www.jstor.org/stable/209516)

Andrade HB, de los Reyes Lopez E, Martin TB (2009) Dimensions of scientific collaboration and its contribution to the academic research groups' scientific quality. Res Eval 18(4):301–311 (DOI: 10.3152/095820209X451041)

Babchuk N, Keith B, Peters G (1999) Collaboration in sociology and other scientific disciplines: A comparative trend analysis of scholarship in the social, physical, and mathematical sciences. Am Sociol 30(3):5–21 (DOI: 10.1007/s12108-999-1007-5, stable JSTOR URL: http://www.jstor.org/stable/27698918)

Barabási AL (2002) Linked: The new science of networks. Perseus, Cambridge, MA

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512 (DOI: 10.1126/science.286.5439.509), also available as arXiv preprint arXiv:cond-mat/9910332

Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311(3–4):590–614 (DOI: 10.1016/S0378-4371(02)00736-7), also available as arXiv preprint arXiv:cond-mat/0104162

Batagelj V, Mrvar A (2010) Pajek – analysis and visualization of large networks: Reference manual: List of commands with short explanations: version 2.00, 1st edn. University of Ljubljana, Ljubljana, available online at the URL: http://vlado.fmf.uni-lj.si/pub/networks/pajek/doc/pajekman.pdf

Batagelj V, Doreian P, Ferligoj A (1992a) An optimizational approach to regular equivalence. Soc Networks 14(1–2):121–135 (DOI: 10.1016/0378-8733(92)90016-Z)

Batagelj V, Ferligoj A, Doreian P (1992b) Direct and indirect methods for structural equivalence. Soc Networks 14(1–2):63–90 (DOI: 10.1016/0378-8733(92)90014-X)

Batagelj V, Ferligoj A, Doreian P (1998) Fitting pre-specified blockmodels. In: Hayashi N, Jajima K, Bock HH, Ohsumi N, Tanaka Y, Baba Y (eds) Data science, classification, and related methods: Proceedings of the fifth conference of the International Federation of Classification

Societies (IFCS-96), Kobe, Japan, March 27–30, 1996. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Tokyo, pp 199–206

Bayer AE, Smart JC (1991) Career publication patterns and collaborative "styles" in american academic science. J High Educ 62(6):613–636 (DOI: http://dx.doi.org/10.2307/1982193, stable JSTOR URL: http://www.jstor.org/stable/1982193)

Berg C (2005) Vernetzung als Syndrom. Risiken und Chancen von Vernetzungsprozessen für eine nachhaltige Entwicklung. Campus: Forschung, vol 883. Campus Verlag, Frankfurt am Main

Bonacich P (2004) The invasion of the physicists. Soc Networks 26(3):285–288 (DOI: 10.1016/j.socnet.2004.06.002)

Bordons M, Morillo F, Gómez I (2004) Analysis of cross-disciplinary research through bibliometric tools. In: Moed HF, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research: The use of patent and publication statistics in studies on S&T systems. Kluwer, Dordrecht, pp 437–456 (DOI: 10.1007/1-4020-2755-9_20)

Braun T, Schubert A (2003) A quantitative view on the coming of age of interdisciplinarity in the sciences 1980–1999. Scientometrics 58(1):183–189 (DOI: 10.1023/A:1025439910278)

Braun-Munzinger P (2009) Teilchenphysik und Teilgebiete der Kernphysik: Das Publikationsverhalten in Großkollaborationen. In: Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen: Beiträge zur Beurteilung von Forschungsleistungen, 2nd ext. edn. Diskussionspapiere der Alexander von Humboldt-Stiftung, No 12. Alexander von Humboldt Foundation/Stiftung, Bonn, pp 92–94, available online at the URL: http://www.humboldt-foundation.de/pls/web/docs/F13905/12_disk_papier_publikationsverhalten2_kompr.pdf

Bucchi M (2004) Science in society: An introduction to social studies of science. Routledge, London. Revised and expanded ed. of: Bucchi M (2002) Scienza e Società: Introduzione alla sociologia della scienza. Il Mulino, Bologna

Burt RS (2004) Structural holes and good ideas. Am J Sociol 110(2):349–399 (DOI: 10.1086/421787, stable JSTOR URL: http://www.jstor.org/stable/10.1086/421787)

Buter RK, Noyons, van Raan AFJ (2010) Identification of converging research areas using publication and citation data. Res Eval 19(1):19–27 (DOI: 10.3152/095820210X492503)

Cole S, Cole JR (1973) Social stratification in science. University of Chicago Press, Chicago, IL

Coleman JS (1964) An introduction to mathematical sociology. Free Press, New York, NY; Collier-Macmillan, London

Crane D (1972) Invisible colleges: Diffusion of knowledge in scientific communities. University of Chicago Press, Chicago, IL

Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a small world from local interactions: Modeling acquaintance networks. Phys Rev Letters 88(12):128701, 4 pp (DOI: 10.1103/PhysRevLett.88.128701), also available as arXiv preprint arXiv:cond-mat/0108302

de Haan J (1997) Authorship patterns in Dutch sociology. Scientometrics 39(2):197–208 (DOI: 10.1007/BF02457448)

de Sola Pool I, Kochen M (1978) Contacts and influence. Soc Networks 1(1):5–51 (DOI: 10.1016/0378-8733(78)90011-4)

Doreian P (2006) Actor network utilities and network evolution. Soc Networks 28(2):137–164 (DOI: 10.1016/j.socnet.2005.05.002)

Doreian P, Stokman FN (eds) (1997) Evolution of social networks. Gordon and Breach, Amsterdam; Routledge, London

Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodeling. Structural Analysis in the Social Sciences, vol 25. Cambridge University Press, Cambridge (DOI: 10.1017/CBO9780511584176)

Erdös P, Rényi A (1959) On random graphs I. Publicationes Mathematicae (Debrecen) 6:290–297, available online at the site: http://www.renyi.hu/~p_erdos/Erdos.html at the URL: http://www.renyi.hu/~p_erdos/1959-11.pdf. Reprinted in: Turán P (ed) (1976) Rényi Alfréd Válogatott Munkái/Selected Papers of Alfréd Rényi, (volume) 2, 1956–1961. Akadémiai Kiadó, Budapest, pp 308–315

Erten C, Harding PJ, Kobourov SG, Wampler K, Yee G (2004) GraphAEL: Graph animations with evolving layouts. In: Liotta G (ed) Graph drawing: 11th international symposium, GD 2003, Perugia, Italy, September 21–24, 2003: Revised papers. Lecture Notes in Computer Science, vol 2912. Springer, Berlin, pp 98–110 (DOI: 10.1007/978-3-540-24595-7_9)

Etzkowitz H, Leydesdorff L (2001) Universities in the global knowledge economy. In: Etzkowitz H, Leydesdorff L (eds) Universities and the global knowledge economy: A triple helix of university-industry-government relations. Science, Technology and the International Political Economy. Pinter, London, pp 1–11

Fienberg SE, Wasserman SS (1981) Categorical data analysis of single sociometric relations. In: Leinhardt S (ed) Sociological methodology 1981. Jossey-Brass, San Francisco, CA, pp 156–192 (DOI: 10.2307/270741 stable JSTOR URL: http://www.jstor.org/stable/270741).

Fienberg SE, Meyer MM, Wasserman S (1985) Statistical analysis of multiple sociometric relations. J Am Stat Assoc 80(389):51–67 (DOI: 10.2307/2288040, stable JSTOR URL: http://www.jstor.org/stable/2288040)

Frank O (1991) Statistical analysis of change in networks. Statistica Neederlandica 45(3):283–293 (DOI: 10.1111/j.1467-9574.1991.tb01310.x)

Frank O, Strauss D (1986) Markov graphs. J Am Stat Assoc 81(395):832–842 (DOI: 10.2307/2289017, stable JSTOR URL: http://www.jstor.org/stable/2289017)

Freeman LC (2004) The development of social network analysis: A study in the sociology of science. EP Empirical Press, Vancouver, BC; BookSurge, North Charleston, SC

Fuchs S (1992) The professional quest for truth: A social theory of science and knowledge. SUNY Series in Science, Technology, and Society. State University of New York Press, Albany, NY

Gansner ER, Koren Y, North SC (2005) Graph drawing by stress majorization. In: Pach J (ed) Graph drawing: 12th international symposium, GD 2004, New York, NY, USA, September 29–October 2, 2004: Revised selected papers. Lecture Notes in Computer Science, vol 3383. Springer, Berlin, pp 239–250 (DOI: 10.1007/978-3-540-31843-9_25)

Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M (1994) The new production of knowledge: The dynamics of science and research in contemporary societies. Sage Publications, London

Glänzel W, Schubert A (2004) Analysing scientific networks through co-authorship. In: Moed HF, Glänzel W, Schmoch U (eds) Handbook of quantitative science and technology research: The use of patent and publication statistics in studies on S&T systems. Kluwer, Dordrecht, pp 257–276 (DOI: 10.1007/1-4020-2755-9_20)

Glänzel W, Schubert A (2005) Domesticity and internationality in co-authorship, references and citations. Scientometrics 65(3):323–342 (DOI: 10.1007/s11192-005-0277-0)

Glänzel W, Schubert A, Czerwon HJ (1999) A bibliometric analysis of international scientific cooperation of the European Union (1985–1995). Scientometrics 45(2):185–202 (DOI: 10.1007/BF02458432)

Gossart C, Ozman M (2009) Co-authorship networks in social sciences: The case of Turkey. Scientometrics 78(2):323–345 (DOI: 10.1007/s11192-007-1963-x)

Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380 (DOI: http://dx.doi.org/10.2307/202051, stable JSTOR URL: http://www.jstor.org/stable/2776392)

Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M (2003) Statnet: Software tools for the statistical modeling of network data, version 2.0. University of Washington, Center for Studies in Demography and Ecology, Seattle, WA, available online at the URL: http://statnetproject.org

Handcock MS, Raftery AE, Tantrum JM (2007) Model-based clustering for social networks. J R Stat Soc Ser A Stat Soc 170(2):301–354 (DOI: 10.1111/j.1467-985X.2007.00471.x), also available as University of Washington, Center for Statistics and the Social Sciences Working Paper no. 46 at the URL: http://www.csss.washington.edu/Papers/wp46.pdf

Havemann F, Heinz M, Kretschmer H (2006) Collaboration and distances between German immunological institutes – a trend analysis. J Biomed Discov Collab 1(6), 7 pp (DOI: 10.1186/1747-5333-1-6)

Hoekman J, Koen F, Tijssen R (2010) Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. Res Policy 39(5):662–673 (DOI: 10.1016/j.respol.2010.01.012)

Holland PW, Leinhardt S (1975) The statistical analysis of local structure in social networks. In: Heise DR (ed) Sociological methodology 1976. Jossey-Brass, San Francisco, CA, pp 1–45 (DOI: 10.2307/270703, stable JSTOR URL: http://www.jstor.org/stable/270703).

Holland PW, Leinhardt S (1977) A dynamic model for social networks. J Math Sociol 5(1):5–20 (DOI: 10.1080/0022250X.1977.9989862)

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76(373):33–50 (DOI: 10.2307/2287037, stable JSTOR URL: http://www.jstor.org/stable/2287037)

Hornbostel S (1997) Bewertungen in der Wissenschaft. Wissenschaftsindikatoren. Westdeutscher Verlag, Opladen

Hummon NP (2000) Utility and dynamic social networks. Soc Networks 22(3):221–249 (DOI: 10.1016/S0378-8733(00)00024-1)

Hummon NP, Carley K (1993) Social networks as normal science. Soc Networks 15(1):71–106 (DOI: 10.1016/0378-8733(93)90022-D)

Hummon NP, Doreian P (1989) Connectivity in a citation network: The development of DNA theory. Soc Networks 11(1):39–63 (DOI: 10.1016/0378-8733(89)90017-8)

Hummon NP, Doreian P (1990) Computational methods for social network analysis. Soc Networks 12(4):273–288 (DOI: 0378-8733(90)90011-W)

Jackson MO, Wolinsky A (1996) A strategic model of social and economic networks. J Econ Theory 71(1):44–74 (DOI: 10.1006/jeth.1996.0108)

Jin EM, Girvan M, Newman MEJ (2001) Structure of growing social networks. Phys Rev E 64(4):046132, 8 pp (DOI: 10.1103/PhysRevE.64.046132)

Katz JS (1994) Geographical proximity and scientific collaboration. Scientometrics 31(1):31–43 (DOI: 10.1007/BF02018100)

Katz JS, Martin BR (1997) What is research collaboration? Res Policy 26(1):1–18 (DOI: 10.1016/S0048-7333(96)00917-1)

Katz L, Proctor C (1959) The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process. Psychometrika 24(4):317–327 (DOI: 10.1007/BF02289814)

Kejžar N (2007) Evolution of networks. PhD thesis. University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana

Kronegger L, Ferligoj A, Doreian P (2011a) On the dynamics of national scientific systems. Quality and Quantity (online first) (DOI: 10.1007/s11135-011-9484-3), in press AQ4

Kronegger L, Mali F, Ferligoj A, Doreian P (2011b) Modeling co-authorship network dynamics among Slovenian researchers. Submitted

Kuhn TS (1996) The structure of scientific revolutions, 3rd edn. University of Chicago Press, Chicago, IL

Laband DN, Tollison RD (2000) Intellectual collaboration. J Polit Econ 108(3):632–662 (DOI: 10.1086/262132, stable JSTOR URL: http://www.jstor.org/stable/10.1086/262132)

Lambiotte R, Panzarasa P (2009) Communities, knowledge creation, and information diffusion. J Informet 3(3):180–190 (DOI: 10.1016/j.joi.2009.03.007), also available as arXiv preprint arXiv:0909.2481

Latapy M, Magnien C, Vecchio ND (2008) Basic notions for the analysis of large two-mode networks. Soc Networks 30(1):31–48 (DOI: 10.1016/j.socnet.2007.04.006)

Laudel G (2002) What do we measure by co-authorship? Res Eval 11(1):3–15 (DOI: 10.3152/147154402781776961)

Leenders RTAJ (1995) Models for network dynamics: A Markovian framework. J Math Sociol 20(1):1–21 (DOI: 10.1080/0022250X.1995.9990149)

Leydesdorff L (2010) Eugene Garfield and algorithmic historiography: Co-words, co-authors, and journal names. Ann Libr Inf Stud 57(3):248–260, available online at the URL: http://nopr.niscair.res.in/handle/123456789/10246, also available as arXiv preprint arXiv:1005.5444. The article is part of: Arunachalam S (ed) Special Issue: A tribute to Eugene Garfield, Information Scientist Extraordinaire, on his 85th birthday. Ann Libr Inf Stud 57(3):173–326

Leydesdorff L, Schank T (2008) Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. J Am Soc Inf Sci Technol 59(11):1532–2890 (DOI: 10.1002/asi.20891), also available as arXiv preprint arXiv:0911.1437

Lorrain F, White HC (1971) Structural equivalence of individuals in social networks. J Math Sociol 1(1):49–80 (DOI: 10.1080/0022250X.1971.9989788)

Lowrie A, McKnight PJ (2004) Academic research networks: A key to enhancing scholarly standing. Eur Manag J 22(4):345–360 (DOI: 10.1016/j.emj.2004.06.011)

Mali F, Kronegger L, Ferligoj A (2010) Co-authorship trends and collaboration patterns in the Slovenian sociological community. Corvinus J Sociol Soc Policy 1(2):29–50, available online at the URL: http://cjssp.uni-corvinus.hu/index.php/cjssp/article/view/24

Marsili M, Vega-Redondo F, Slanina F (2004) The rise and fall of a networked society: A formal model. Proc Natl Acad Sci USA 101(6):1439–1442 (DOI: 10.1073/pnas.0305684101, stable JSTOR URL: http://www.jstor.org/stable/3371207)

Matthew D (2005) Science in society. Palgrave Macmillan, Basingstoke

Mayer T (1984) Parties and networks: Stochastic methods for relationship networks. J Math Sociol 10(1):51–103 (DOI: 10.1080/0022250X.1984.9989958)

Melin G, Persson O (1996) Studying research collaboration using co-authorship. Scientometrics 36(3):363–377 (DOI: 10.1007/BF02129600)

Merton RK (1973) The sociology of science: Theoretical and empirical investigations. University of Chicago Press, Chicago, IL

Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Structures Algorithms 6(2–3):161–179 (DOI: 10.1002/rsa.3240060204)

Moody J (2004) The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. Am Sociol Rev 69(2):213–238 (DOI: 10.1177/000312240406900204, stable JSTOR URL: http://www.jstor.org/stable/3593085)

Mulkay M (1975) Three models of scientific development. Sociol Rev 23(3):509–526

Munshi UM, Pant P (2004) Regional, national, international coauthorship and subauthorship collaboration and author productivity as manifested in the scholarly journal literature of nuclear science. In: Kretschmer H, Singh Y, Kundra R (eds) International Workshop on Webometrics, Informetrics and Scientometrics (WIS 2004), Roorkee, March 2–5, 2004. Indian Institute of Technology, Roorkee, pp 273–290, available online at the URL: http://hdl.handle.net/123456789/461

Newman MEJ (2000) Models of the small world. J Stat Phys 101(3–4):819–841 (DOI: 10.1023/A:1026485807148), also available as arXiv preprint arXiv:cond-mat/0001118

Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98(2):404–409 (DOI: 10.1073/pnas.021544898, stable JSTOR URL: http://www.jstor.org/stable/3054694), also available as arXiv preprint arXiv:cond-mat/0007214

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256 (DOI: 10.1137/S003614450342480), also available as arXiv preprint arXiv:cond-mat/0303516

Newman MEJ (2004) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci USA 101(Suppl 1):5200–5205 (DOI: 10.1073/pnas.0307545100). The article is part of the special issue: (2004) Mapping knowledge domains: Arthur M. Sackler Colloquium of the National Academy of Sciences. Proc Natl Acad Sci USA 101(Suppl 1):5183–5310

Nooraie RY, Kamrani MA, Hanneman R, Etemadi A (2008) Association between co-authorship network and scientific productivity and impact indicators in academic medical research centers: A case study in Iran. Health Res Policy Syst 6(9), 8 pp (DOI: 10.1186/1478-4505-6-9)

Opsahl T (2010) Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. Arxiv preprint arXiv:1006.0887

Pattison P, Wasserman S (1999) Logit models and logistic regressions for social networks: II. Multivariate relations. Br J Math Stat Psychol 52(2):169–193 (DOI: 10.1348/000711099159053)

Pennock DM, Flake GW, Lawrence S, Glover EJ, Giles CL (2002) Winners don't take all: Characterizing the competition for links on the web. Proc Natl Acad Sci USA 99(8):5207–5211 (DOI: 10.1073/pnas.032085699, stable JSTOR URL: http://www.jstor.org/stable/3058470)

Perc M (2010) Growth and structure of Slovenia's scientific collaboration network. J Informetr 4(4):475–482 (DOI: 10.1016/j.joi.2010.04.003), also available as arXiv preprint arXiv:1004.4824

Powell WW, White DR, Koput KW, Owen-Smith J (2005) Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. Am J Sociol 110(4):1132–1205 (DOI: 10.1086/421508, stable JSTOR URL: http://www.jstor.org/stable/10.1086/421508)

Price DJ de Solla (1963) Little science, big science. Columbia University Press, New York, NY

Price DJ de Solla (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306 (DOI: 10.1002/asi.4630270505), also available online at the URL: http://www.asis.org/Publications/JASIS/Best_Jasist/1976pricejasistarticle.pdf

Price DJ de Solla, Beaver D (1966) Collaboration in an invisible college. Am Psychologist 21(11):1011–1018 (DOI: 10.1037/h0024051), also available online at the URL: http://www.sciencedirect.com/science/article/pii/S0003066X07631281

Qin J, Lancaster FW, Allen B (1997) Types and levels of collaboration in interdisciplinary research in the sciences. J Am Soc Inf Sci 48(10):893–916 (DOI: 10.1002/(SICI)1097-4571(199710)48:10¡893::AID-ASI5¿3.0.CO;2-X)

Richardson MW (1938) Multidimensional psychophysics. Psychol Bull 35:659–660

Rigby J (2005) Handcrafted by 16 men: The impact of single and multiple authorship in collaborative research networks. Res Eval 14(3):199–206 (DOI: 10.3152/147154405781776148)

Rogers J, Bozeman B, Chompalov I (2001) Obstacles and opportunities in the application of network analysis to the evaluation on R&D. Res Eval 10(3):161–172 (DOI: 10.3152/147154401781777033)

Said YH, Wegman EJ, Sharabati WK, Rigsby JT (2008) Social networks of author-coauthor relationships. Comput Stat Data Anal 52(4):2177–2184 (DOI: 10.1016/j.csda.2007.07.021)

Sailer L (1978) Structural equivalence: Meaning and definition, computation and application. Soc Networks 1(1):73–90 (DOI: 10.1016/0378-8733(78)90014-X)

Schott T (1993) World science: Globalization of institutions and participation. Sci Technol Human Values 18(2):196–208 (DOI: 10.1177/016224399301800205, stable JSTOR URL: http://www.jstor.org/stable/690067)

Sigogneau A, Malagutti O, Crance M, Bauin S (2005) Cross-disciplinary research: Co-evaluation and co-publication practices of the CNRS laboratories. Res Eval 14(2):165–176 (DOI: 10.3152/147154405781776210)

Skyrms B, Pemantle R (2000) A dynamic model of social network formation. Proc Natl Acad Sci USA 97(16):9340–9346 (DOI: 10.1073/pnas.97.16.9340, stable JSTOR URL: http://www.jstor.org/stable/123177)

Snijders TAB (1990) Testing for change in a digraph at two time points. Soc Networks 12(4):359–373 (DOI: 10.1016/0378-8733(90)90015-2)

Snijders TAB (1995) Methods for longitudinal social network data. In: Tiit EM, Kollo T, Niemi H (eds) Multivariate statistics and matrices in statistics: Proceedings of the 5th Tartu Conference, Tartu, Pühajärve, Estonia, 23–28 May 1994. New Trends in Probability and Statistics, vol 3. TEV, Vilnius; VSP, Utrecht, pp 221–227

Snijders TAB (1996) Stochastic actor-oriented models for network change. J Math Sociol 21(1&2):149–176 (DOI: 10.1080/0022250X.1996.9990178)

Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology: Volume 31: 2001. Blackwell, Boston, MA, pp 361–395 (DOI: 10.1111/0081-1750.00099, stable JSTOR URL: http://www.jstor.org/stable/3097281)

Snijders TAB, Steglich CEG, Schweinberger M, Huisman K (2008) Manual for SIENA version 3.2. University of Groningen, ICS, Department of Sociology, Groningen; University of Oxford, Department of Statistics, Oxford, available online at the URL: http://stat.gamma.rug.nl/s_man317.pdf and at the URL: http://www.stats.ox.ac.uk/ snijders/siena/s_man312.pdf

Snijders TAB, van de Bunt GG, Steglich CEG (2010) Introduction to stochastic actor-based models for network dynamics. Soc Networks 32(1):44–60 (DOI: 10.1016/j.socnet.2009.02.004)

Sonnenwald DH (2007) Scientific collaboration. Ann Rev Inf Sc Technol 41:643–681 (DOI: 10.1002/aris.2007.1440410121)

Tomassini M, Luthi L (2007) Empirical analysis of the evolution of a scientific collaboration network. Physica A 385(2):750–764 (DOI: 10.1016/j.physa.2007.07.028)

Torgerson WS (1952) Multidimensional scaling: I. Theory and method. Psychometrika 17(4):401–419 (DOI: 10.1007/BF02288916)

Travers J, Milgram S (1969) An experimental study of the small world problem. Sociometry 32(4):425–443 (DOI: 10.2307/2786545, stable JSTOR URL: http://www.jstor.org/stable/2786545)

Vázquez A (2003) Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Phys Rev E 67(5):056104, 15 pp (DOI: 10.1103/PhysRevE.67.056104), also available as arXiv preprint arXiv:cond-mat/0211528

Veugelers R, Cassiman B (2005) R&D cooperation between firms and universities: Some empirical evidence from Belgian manufacturing. Int J Ind Organization 23(5–6):355–379 (DOI: 10.1016/j.ijindorg.2005.01.008), also available online as Katholieke Universiteit Leuven, Department Toegepaste Economische Wetenschappen Research Report 0325 at the URL: https://lirias.kuleuven.be/bitstream/123456789/118346/1/OR_0325.pdf

Wagner CS (2005) Six case studies of international collaboration in science. Scientometrics 62(1):3–26 (DOI: 10.1007/s11192-005-0001-0)

Wagner CS, Leydesdorff L (2005) Network structure, self-organization, and the growth of international collaboration in science. Res Policy 34(10):1608–1618 (DOI: 10.1016/j.respol.2005.08.002), also available as arXiv preprint arXiv:0911.4299

Wang P, Robins G, Pattison P (2009) PNet: Program for the simulation and estimation of ($p^{\star}$) exponential random graph models: User Manual. University of Melbourne, Department of Psychology, Melbourne, available online at the URL: http://www.sna.unimelb.edu.au/pnet/pnet.html

Wasserman SS (1978) Models for binary directed graphs and their applications. Adv Appl Probab 10(4):803–818 (DOI: 10.2307/1426660, stable JSTOR URL: http://www.jstor.org/stable/1426660)

Wasserman S (1980a) Analyzing social networks as stochastic processes. J Am Stat Assoc 75(370):280–294 (DOI: 10.2307/2287447, stable JSTOR URL: http://www.jstor.org/stable/2287447)

Wasserman SS (1980b) A stochastic model for directed graphs with transition rates determined by reciprocity. In: Schuessler KF (ed) Sociological methodology 1980. Jossey-Bass, San Francisco, CA, pp 392–412 (DOI: 10.2307/270870, stable JSTOR URL: http://www.jstor.org/stable/270870).

Wasserman S (1987) Conformity of two sociometric relations. Psychometrika 52(1):3–18 (DOI: 10.1007/BF02293953)

Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Structural Analysis in the Social Sciences, vol 8. Cambridge University Press, Cambridge

Wasserman S, Iacobucci D (1988) Sequential social network data. Psychometrika 53(2):261–282 (DOI: 10.1007/BF02294137)

Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. Psychometrika 61(3):401–425 (DOI: 10.1007/BF02294547)

Wasserman S, Weaver SO (1985) Statistical analysis of binary relational data: Parameter estimation. J Math Psychol 29(4):406–427 (DOI: 10.1016/0022-2496(85)90003-3)

Watts DJ (2003) Six degrees: The science of a connected age, 1st edn. W. W. Norton & Company, New York, NY

Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. Nature 393(6684):440–442 (DOI: 10.1038/30918)

White D, Reitz K (1983) Graph and semigroup homomorphisms on networks of relations. Soc Networks 5(2):193–234 (DOI: 10.1016/0378-8733(83)90025-4)

Whitley R (2000) The intellectual and social organization of the sciences. Oxford University Press, Oxford

Wray KB (2002) The epistemic significance of collaborative research. Philos Sci 69(1):150–168 (DOI: 10.1086/338946, stable JSTOR URL: http://www.jstor.org/stable/3081018)

Xie F, Cui W (2008a) Cost range and the stable network structures. Soc Networks 30(1):100–101 (DOI: 10.1016/j.socnet.2007.07.002)

Xie F, Cui W (2008b) A note on the paper 'cost range and the stable network structures'. Soc Networks 30(1):102–103 (DOI: 10.1016/j.socnet.2007.07.005)

Ziman J (1994) Prometheus bound: Science in dynamic steady state. Cambridge University Press, Cambridge

Ziman J (2000) Real science: What it is, and what it means. Cambridge University Press, Cambridge (DOI: 10.1017/CBO9780511541391)

Zuckerman H (1967) Nobel laureates in science: Patterns of productivity, collaboration, and authorship. Am Sociol Rev 32(3):391–403 (DOI: 10.2307/2091086, stable JSTOR URL: http://www.jstor.org/stable/2091086)

AUTHOR QUERIES

AQ1. Please provide reference list for Börner et al. (2010).
AQ2. Please check whether the inserted citation for Table 6.1 is appropriate.
AQ3. Please check whether the inserted citation for Figs. 6.2 and 6.3 are appropriate.
AQ4. Please update ref. Kronegger et al. (2011a,b)

Chapter 7
Citation Networks

Filippo Radicchi, Santo Fortunato, and Alessandro Vespignani

7.1 Introduction

AQ1 Bibliographic databases represent the starting point for any empirical study of the evolution and dynamics of scientific activity, citation patterns, and the ensuing analysis of the importance of specific contributions, journals, and scientists. Bibliographic datasets were first analyzed by Lotka (1926) and Shockley (1957) in order to quantitatively measure the productivity of individual scientists and research laboratories, respectively. Since the pioneering work of Derek de Solla Price (1965), who realized that bibliographic data have a natural mathematical representation in terms of directed graphs, the study of co-authorship and citation networks has become the starting point for the formulation of key hypotheses such as the mechanism of cumulative advantage (Price 1976) to explain the dynamical pattern of citation accumulation. The mathematical description of social systems in terms of networks or graphs has a long tradition in social sciences (Wasserman and Faust 1994). However, it is only in the last decade that the analysis of bibliographic data has received a boost from advances in information technology

---

F. Radicchi
Howard Hughes Medical Institute (HHMI), Northwestern University, Evanston,
Illinois 60208, USA
e-mail: f.radicchi@gmail.com

S. Fortunato
Complex Networks and Systems, ISI Foundation, Torino, Italy
e-mail: fortunato@isi.it

A. Vespignani
Center for Complex Networks and Systems Research (CNetS), School of Informatics Computing,
Indiana University, Bloomington, Indiana 47408, USA
e-mail: alexv@indiana.edu

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_7,

and the massive digitalization of documents. For the first time, data collection and mining capabilities allow for systems-level analysis of huge bibliometric datasets that are regularly collected in digital format. The data collected in digital bibliographic databases report a wealth of information for each article, including: title, journal, date of publication, a list of authors and their affiliations, a list of bibliographic references, keywords, and an abstract. In this context, the use of multipartite networks as the natural abstract mathematical representation of the data is particularly convenient, and several studies have recently focused on the study of co-authorship networks, paper citation networks, etc. In general, each of these networks is an appropriate bipartite or unipartite network projection of the original bibliographic dataset where authors and papers are nodes, and citations, authorship, and other bibliographic information define the links between nodes.

Nowadays, computational power allows us to generate and analyze citation networks consisting of hundreds of thousands or millions of nodes and links. On one hand, the sheer size of the networks under consideration challenges us with new problems concerning the mathematical characterization of systems that preserve the undeniable intricacies and, in some cases, haphazard sets of elements and relations involved. On the other hand, the large size of the resulting networks empowers us with a systems-level view of the citation dynamics that was not accessible in previous years. Indeed, in large systems, asymptotic regularities cannot be detected by looking at local elements or properties: one has to shift attention to statistical measures that take into account the global behavior of these quantities.

The possibility of analyzing large-scale network data is one of the central elements that has characterized the recent developments in network science and the increased interest in complex networks (Albert and Barabási 2002; Dorogovtsev and Mendes 2002; Newman 2003; Pastor-Satorras and Vespignani 2004; Boccaletti et al. 2006; Caldarelli 2007; Barrat et al. 2008). For this reason, citation networks in the last several years have become one of the prototypical examples of complex network evolution. Indeed, the new modeling and analysis techniques emerging in the area of complex networks have provided new insights into citation networks, which have facilitated understanding of the dynamical processes governing their evolution. In this chapter, we will review the main structural characteristics of citation networks and we will frame some of their properties in the language of complex networks. We will also review the basic descriptive and generative models used to represent citation networks and the use of dynamical processes to rank papers and authors (Table 7.1). AQ2

Table 7.1 List of major questions and models addressed in this chapter

Major questions addressed	Major models
Structure and dynamics of citation networks	De Solla Price model
Citation distributions	Barabási-*Albert model
Ranking criteria	Model by Karrer & Newman

7.2 Bibliographic Databases and the Construction of Citation Networks

In the last two decades, bibliographic databases have completely changed in terms of accessibility and completeness. Most of these databases are now online and their records can be searched by simple web queries. The *Web of Science* (WoS) database of Thomson Reuters[1] is the largest and most complete commercial source of bibliographic data. WoS indexes papers from every part of the world and from every scientific discipline. Like WoS, other databases store large sets of bibliographic data: *CrossRef*,[2] *Scopus*,[3] *GoogleScholar*,[4] *Citebase*,[5] *CiteSeer*,[6] *Spires*,[7] and the Eprint archive at `www.arxiv.org` are just a few examples. These databases do not offer the same coverage of WoS (different journals and conference proceedings are listed depending on the database), but, with the exception of CrossRef and Scopus, they are accessible free of charge.

From the raw data, various kinds of citation graphs can be generated. The simplest ones are citation networks between papers. Taking the list of references appearing at the end of each article, one can draw directed connections from citing articles to cited ones. In this case, the graph is directed, but no weight appears on the arcs since it is natural to assume that each reference has the same importance. The same information can be used to construct citation networks between scientists, journals, and institutions. For example, the citation network between journals is obtained by substituting each article with its journal of publication. Weighted connections can be drawn in this case by assigning to the arcs a weight equal to the number of times that a journal cites another journal. In Fig. 7.1, we show the construction of an author citation network. Starting from the network of citations between papers, the construction can be performed locally by translating the citation from a paper i to a paper j into a set of citations between all n_i co-authors of paper i to all n_j co-authors of article j. The weight of each of these directed connections is simply $w = 1/\left(n_i \cdot n_j\right)$, by naturally assuming that the citation between papers carries a unit of weight and that this quantity is evenly split among the involved scientists. The total weight of a connection between two authors is then given by the sum of each of these elementary contributions over the entire network of citations between papers. Furthermore, the longitudinal nature of bibliographic datasets (expressed by the publication dates of the papers) allows one to follow the evolution of citation networks.

---

[1] WoS: Web of Science, URL: http://isiknowledge.com/WOS.

[2] Crossref, URL: http://www.crossref.org.

[3] Scopus, URL: http://www.scopus.com.

[4] Google Scholar, URL: http://scholar.google.com.

[5] Citebase, URL: http://www.citebase.org.

[6] Citeseer, URL: http://citeseer.ist.psu.edu.

[7] SPIRES, URL: http://www.slac.stanford.edu/spire.

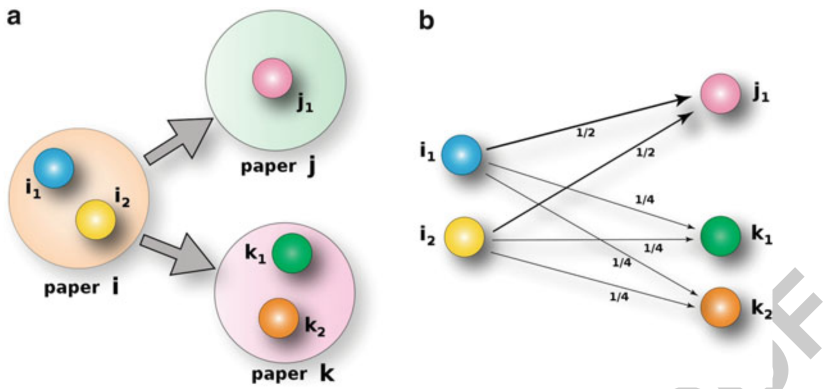


Fig. 7.1 (**a**) In the network of citations between papers, the article i, written by two authors i_1 and i_2, cites two papers j and k, written by one author j_1 and two co-authors k_1 and k_2, respectively. (**b**) The author citation projection is generated by simply connecting with a directed link both i_1 and i_2 to j_1, each with weight $1/2$, and to k_1 and k_2, each with weight $1/4$. From Radicchi et al. (2009)

It is important to mention that in the construction of citation (and collaboration) networks between scientists, possible problems may arise. First, there is a problem of identification for the authors. Unfortunately, scientists do not always sign their papers using the same name and this has as a consequence the impossibility of automatically relating different names to the same physical person. This fact may happen for several reasons: different order between first and last name; possible presence or absence of middle names; and change of last names (especially after marriage). The second problem is basically the reverse of the formerly described source of error. Generally in bibliographic databases, scientists are identified by their full last name plus the initials of their first and middle names. Therefore, disambiguation errors occur due to the impossibility of distinguish authors having the same initials and the same last name. The solution for deleting these source of errors is to use a unique identifier for each scientist as recently proposed by the project ResearcherID[8] of Thomson Scientific.

It is worth remarking that citation networks can also be constructed by considering data, and not concerning the scientific bibliography. For instance, there is a large number of electronic databases collecting information on technological patents. Examples are: NBER U.S. Patent Citations Data,[9] containing all patents registered in the United States from 1963 to 1999; Google patents,[10] which collects patents

---

[8] URL: http://www.researcherid.com.

[9] NBER: The National Bureau of Economic Research, U.S. Patent Citations Data at the URL: http://www.nber.org/patents.

[10] Google Patents, URL: http://www.google.com/patents.

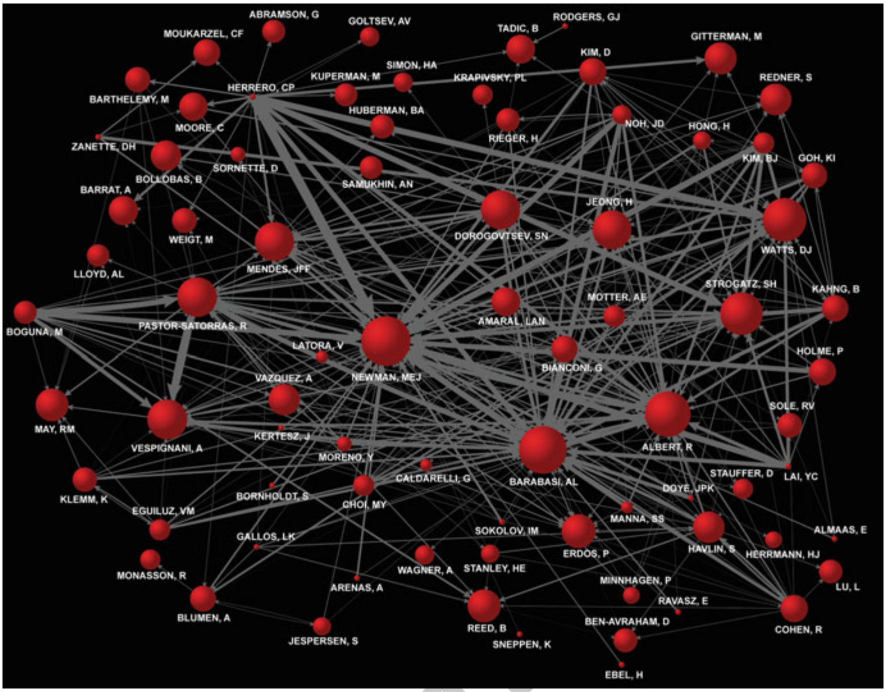


Fig. 7.2 Author citation network of scientists working on complex networks. The graph is derived from the citation network of papers published in the journals of the American Physical Society[14] whose titles contain the keywords "complex networks," "small-world networks," etc. A citation by paper A to paper B turns into a set of citations from each author of paper A to each author of paper B. Each edge of the author citation network is weighted, as an author may cite any other author multiple times in the same or different papers. From Radicchi et al. (2009)

registered in many countries; and the database of the European Patent Office,[11] in which all patents registered in the European Community are stored. An additional example is represented by legal citation networks. These are networks that can be constructed by using data obtained from United States Supreme Court decisions dating from 1789.[12]

Citation networks (see Fig. 7.2) immediately convey a sense of complexity, and, in order to understand the organizing principles underlying these networks, it is necessary to utilize statistical analysis. The first quantity to be scrutinized since the early work of De Solla Price has been degree centrality. The degree k_i of a vertex i is defined as the number of edges in the graph incident on the vertex i. While this definition is clear for undirected graphs, it needs some refinement for the case of directed graphs. Thus, we define the *in-degree* k_i^{in} of the vertex i as the number

---

[11] EPO: European Patent Office, URL: http://www.epo.org/patents/patent-information.html.

[12] Supreme Court of the United States, URL: http://www.supremecourt.gov.

of edges arriving at i, while its *out-degree* k_i^{out} is defined as the number of edges departing from i. The degree of a vertex in a directed graph is defined by the sum of the in-degree and the out-degree, $k_i = k_i^{in} + k_i^{out}$. In the case of paper citation networks, the in-degree k_i^{in} corresponds to the number of papers citing the paper i and the out-degree k_i^{out} corresponds to the number of citations to other papers. In large-scale graphs, a first statistical characterization is provided by the normalized histogram of the in-degree and out-degree of the nodes that for a large number of nodes (documents) can be considered analogous to the probability distributions $P\left(k^{in}\right)$ and $P\left(k^{out}\right)$ that a randomly chosen vertex has in-degree k^{in} and out-degree k^{out}, respectively. While these two quantities have been considered extensively in the literature, it is clear that many other indicators and metrics characterizing the structure of networks are equally important in defining the ordering principles of citation networks. In the next section, we will discuss some of the structural features that characterize citation networks. However, it is important to stress that the analysis of the degree distributions of citation networks immediately reveals a high level of heterogeneity exemplified by the fact that many vertices have just a few connections, while a few hubs collect hundreds or even thousands of edges. For instance, this feature is easily discerned from Fig. 7.2. The same arrangement can easily be perceived in many other networks where the presence of "hubs" is a natural consequence of different factors such as popularity, strategies, and optimization. For instance, in the World Wide Web, some pages become hugely popular and are pointed to by thousands of other pages, while, in general, the majority of pages are almost unknown. The presence of hubs and connectivity define degree distributions $P(k^{in})$ with heavy-tails (Barabási and Albert 1999) that are highly variable in the sense that degrees vary over a broad range, spanning several orders of magnitude. This behavior is very different from the case of bell-shaped, exponentially decaying distributions. In distributions with heavy tails, vertices with degrees much larger than the average $\langle k^{in} \rangle$ are found with a significant probability. In other words, the average behavior of the system is not typical.

The heterogeneity found in citation networks is common to many other networks in very different domains. This evidence, first pointed out by Barabási and Albert (1999), is at the root of the huge body of work aimed at uncovering general dynamical principles explaining the structure and evolution of complex networks. It is necessary however to clarify the distinction between what is "complex" and what is merely complicated, in addition to what is conceptually relevant to citation networks. A first point which generally characterizes complex systems is that they are emergent phenomena in the sense that they are the spontaneous outcome of the interactions of many constituent units. In other words, complex systems are not engineered systems put in place according to a definite blueprint. Indeed, loosely speaking, complex systems consist of a large number of elements capable of interacting with each other and their environment in order to organize within specific emergent structures. From this perspective, another characteristic of complex systems is that decomposing the system and studying each component in isolation does not allow for an understanding of the whole system and its dynamics since the self-organization principles reside mainly in the collective and unsupervised dynamics of the many elements. It is easy to see that citation networks are this type of systems. Another main feature characterizing many complex systems concerns the presence of complications on all scales possible within the physical constraints of the system. In other words, when facing complex systems, we are in the presence of structures whose fluctuations and heterogeneities extend and are repeated at all scales of the system. In the case of citation networks, the all-scales complication is statistically encoded in the heavy-tail distributions that characterize network structural properties. The larger the size of a system, the larger its heterogeneity and the variability of its properties.

The question of the existence of some general organizing principles that might explain the emergence of complex networks architecture in very different contexts leads naturally to a shift of focus in the area of network modeling where the emphasis is on the microscopic processes that govern the appearance and disappearance of vertices and links. In this context, citation networks have acquired a role that goes beyond the specific interest of bibliometrics and the so-called "science of science"; they are prototypical systems for the study of dynamical principles that could apply in very different domains.

7.3 Structural Features of Citation Networks

7.3.1 Citation Distribution

The primary goal of a large number of empirical studies about citation networks is represented by the characterization of the probability distribution function of citations. This is the probability $P\left(k^{in}\right)$ that a paper has been cited k^{in} times. In the language of network science, measuring the number of citations of a paper means counting the number of incoming links (in-degree) k^{in} of a node. In the 1960s, de Solla Price (Price 1965) was already in the middle of performing empirical measurements on a relatively small subset of papers and was able to observe that the number of articles with a given number of citations had a broad distribution. Price conjectured a power law scaling $P\left(k^{in}\right) \sim \left(k^{in}\right)^{-\gamma}$ with a decaying exponent $\gamma \simeq 3$. This result was confirmed much later in 1998 by Redner (1998). Redner studied much larger datasets (all papers published in *Physical Review D* up to 1997 and all articles indexed by Thomson Scientific in the period from 1981–1997) and found again that the right tail of the distribution (corresponding to highly cited papers) shows a power law scaling with $\gamma = 3$. At the same time, Redner realized that the left part of the distribution was more consistent with a stretched exponential. However, different conclusions were drawn by Laherrére and Sornette (1998) in the same year. They studied the dataset of the top 1,120 most cited physicists during the period from 1981–1997, finding that the whole distribution of citations is more compatible with a stretched exponential $P\left(k^{in}\right) \sim \exp\left[-\left(k^{in}\right)^{\beta}\right]$, with $\beta \simeq 0.3$. The puzzle was seemingly solved by Tsallis and de Albuquerque

(2000). By analyzing the same datasets as Redner's plus an additional one composed of all the papers published up to 1999 in *Physical Review E*, the authors found that the Tsallis distribution $P\left(k^{in}\right) = P(0)/\left[1 + (\beta - 1)\,\lambda\, k^{in}\right]^{\beta/(\beta-1)}$, with $\lambda \simeq 0.1$ and $\beta \simeq 1.5$, consistently fits the entire distribution of citations. However, a new functional form was again attributed to Redner a little later. Redner performed an analysis over all papers published in the 110-years-long history of journals in the *Physical Review* collection (Redner 2005), finding that the distribution of citations is best fitted by a log-normal distribution

$$P\left(k^{in}\right) = \frac{1}{k^{in}\sqrt{2\pi\sigma^2}}\ \exp\left\{-\left[\ln\left(k^{in}\right) - \mu\right]^2 / \left(2\sigma^2\right)\right\}. \tag{7.1}$$

In subsequent studies, depending on the particular dataset taken under consideration, distributions of citations have been fitted with various functional forms: power-laws (Seglen 1999; Vazquez 2001; Lehmann et al. 2003; Bommarito and Katz 2009), log-normals (Bommarito and Katz 2009; Stringer et al. 2008; Radicchi et al. 2008; Castellano and Radicchi 2009; Stringer et al. 2010), Tsallis distributions (Wallace et al. 2009; Anastasiadis et al. 2009), modified Bessel functions (van Raan 2001a,b), and more complicated distributions (Kryssanov et al. 2007).

A typical bias present in many empirical results is the fact that citation distributions are computed without taking into consideration any possible discipline- or age-dependence of the statistics. Older papers may have more citations than recent ones, not necessarily because of their merits, but because they stayed in the literature longer and had more time to be cited. Even more serious is the bias related to discipline dependence: papers in mathematics and biology are part of two almost non-interacting citation networks, which follow different citing behaviors. In Stringer et al. (2008); Radicchi et al. (2008); Castellano and Radicchi (2009); Stringer et al. (2010), the authors accounted for these distinctions by analyzing a large number of papers and classifying them according to the date and the journal of publication (Stringer et al. 2008, 2010) and the scientific discipline to which they belong (Radicchi et al. 2008; Castellano and Radicchi 2009). By restricting the statistic to these subsets, the probability that a paper has received k^{in} citations is a log-normal distribution. Even more surprisingly, the authors of Radicchi et al. (2008) realized that the only significant difference between different disciplines and years of publication is the average value $\langle k^{in}\rangle$. When the raw number of citations is replaced by the relative quantity $k^{in}/\langle k^{in}\rangle$, a universal behavior is found and no distinction between curves corresponding to different publication years and scientific disciplines is visible (Fig. 7.3).

7.3.2 *Other Topological Features of Citation Networks*

Citation networks are directed graphs, and typical measurements used for undirected networks must be adapted. Directions are naturally defined, since the arrows on the arcs of the graph point from citing to cited articles. In good approximation, paper

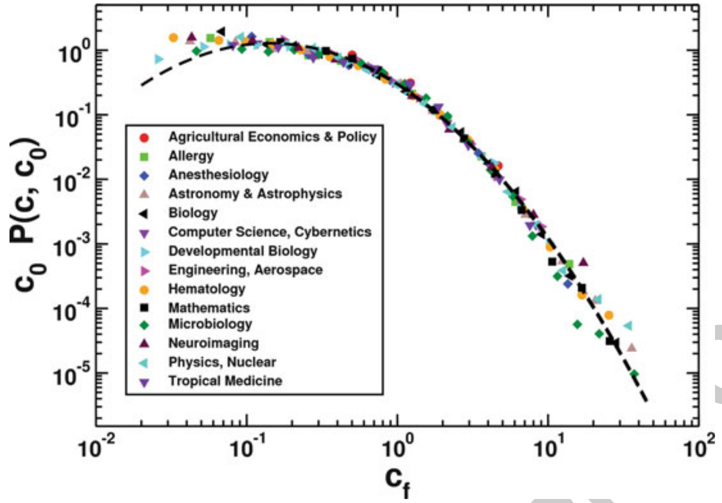


Fig. 7.3 Universality of citation distributions. Each curve refers to papers published in a given year in journals belonging to the same discipline. The disciplines are those identified by ISI Web of Science. The score on the x-axis is the ratio of the number of cites c of a paper by the average number of cites c_0 collected by all papers in that discipline. From Radicchi et al. (2008)

citation networks are also acyclic graphs. The lack of cycles is due to the natural order underlying the network: papers are chronologically sorted, and citations only go backward in time. However, this feature is not generally present in citation networks, as, for example, in citation graphs between scientists and journals. Moreover, though in rare cases, paper citation networks are not strictly acyclic since special issues of journals often contain articles citing each other.

Triangles (important for computing local correlation properties like the clustering coefficient (Watts and Strogatz 1998)) can be still observed, but only of the type $i \to j, i \to l$, and $j \to l$. These local structures abound in scientific citation networks (Chen et al. 2007; Wu and Holme 2009): generally speaking, in 50% of the cases the presence of the citations $i \to j$ and $j \to l$ also implies the existence of the arc $i \to l$. This means that there is a general tendency to copy the references of cited papers. An interesting consequence of this mechanism is the spreading of errors in referenced papers (Simkin and Roychowdhury 2005), due to the fact that often citations are copied from other papers without paying attention to their correctness.

Another general difference with respect to undirected networks is the presence in citation graphs of "sinks": i.e., papers that do not cite any article and have therefore zero out-degree. The presence of sinks is generally due to the incompleteness of the datasets; the oldest papers indeed cite other articles, but those cited articles are not included in the analysis as they are even older than the citing article. Similarity indexes and distances can be formulated despite this. In Bommarito et al. (2010b) for example, the distance between two nodes is quantified in terms of common ancestors (sinks). The degree of similarity can be used for classification purposes through the application of data clustering algorithms.

7.3.3 Community Structure of Citation Networks

Real networks typically display an internal organization of clusters (communities). Communities are intuitively defined as sets of vertices characterized by a density of internal connections higher than the density of links between vertices of different communities. The identification of communities in complex networks is a non-trivial problem, originally considered in social science (Scott 2000) and later analyzed in theoretical computer science in the context of the data clustering problem (Jain et al. 1999). Recently, concepts and tools typical of statistical physics have played a fundamental role for the detection of topological communities in complex networks (Fortunato 2010).

Citation networks represent a difficult challenge for community detection. Since they are directed (sometimes weighted) graphs with an internal natural ordering

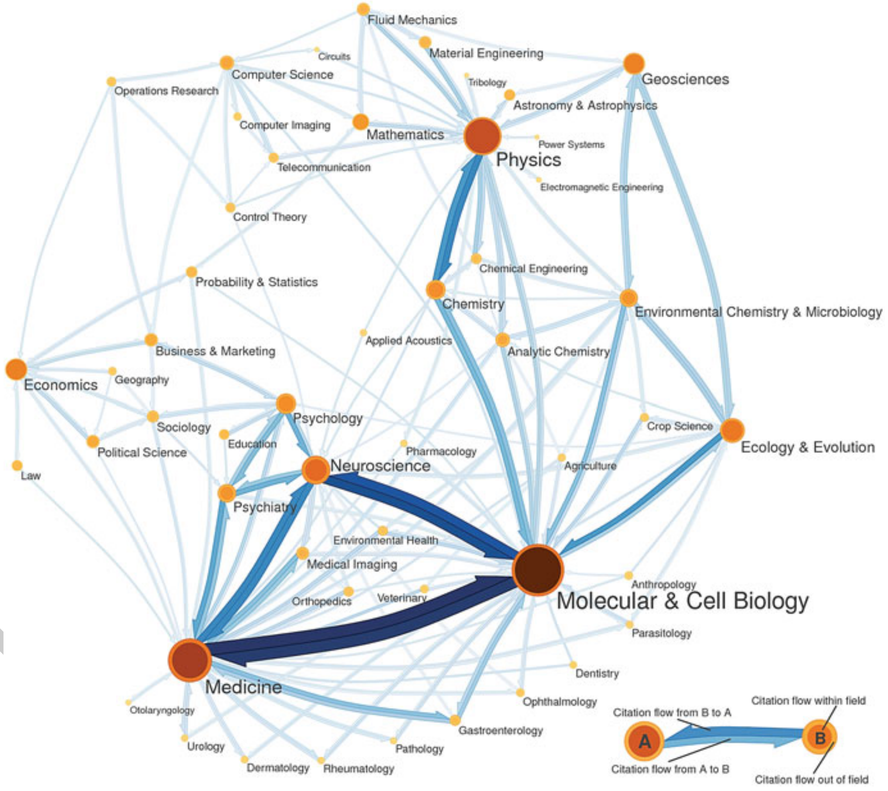


Fig. 7.4 Community structure of a network of scientific journals. Communities, indicated by the *circles*, were detected via Infomap, an algorithm based on the study of diffusion flows in the network. Each *circle* is named after the discipline of the journals grouped in the corresponding community. The thickness of the arcs is proportional to the size of the citation flows between disciplines. From Rosvall and Bergstrom (2008)

(publication time), standard tools of community detection, generally developed for undirected and unweighted graphs, require modification, and in most cases this is not possible. Fortunately, some new techniques for community detection have been developed and applied for the identification of clusters in citation networks.

One interesting approach is the one proposed by Rosvall and Bergstrom (2008). Using an information-theoretic framework, based on coding of diffusion processes on graphs, the authors were able to determine the community structure in the citation network between the scientific journals indexed by Thompson Scientific, identifying the main divisions of journals in scientific disciplines (Fig. 7.4). A different kind of analysis is the one recently performed by Chen and Redner (2010). The authors studied the community structure of the citation network between papers published in the collection of *Physical Review* by means of maximization of the directed version of the modularity function (Leicht and Newman 2008). The study by Chen and Redner leads to the observation of the surprising presence of strong connections between fields of physics that are *prima facie* very different with respect to research topic or that are well-separated in time (Fig. 7.5). Other interesting approaches are those proposed for the study of the community structure of the legal citation network of the Supreme Court of the United States (Leicht et al. 2007; Bommarito et al. 2010a). In Leicht et al. (2007), an expectation-maximization algorithm is used for monitoring the evolution of communities. In Bommarito et al. (2010a), communities are found through different detection algorithms and their stability along time is controlled.

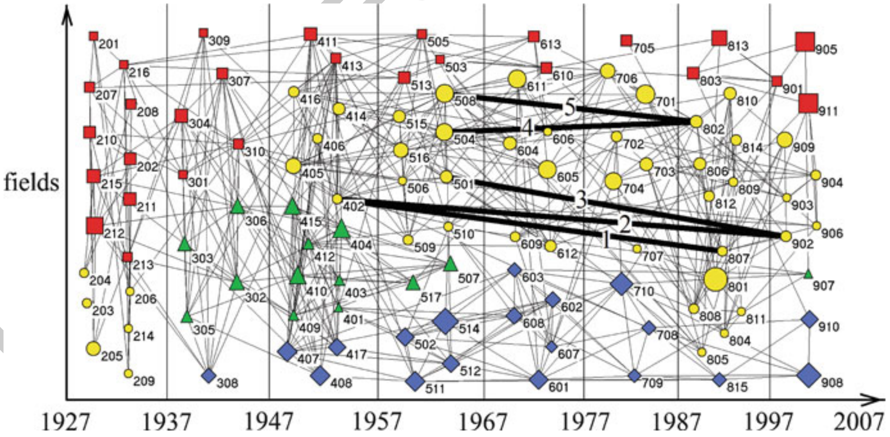


Fig. 7.5 Time evolution of the community structure of the network of citations between papers published in journals of the American Physical Society (APS). Time is divided into nine decades, from 1927 until 2006. In each decade, the most cited papers were selected (about 3,000). The communities are classified based on the APS journal where the largest relative fraction of papers in the community were published (indicated by the symbols). While links between different decades usually involve consecutive periods, there are five links connecting well-separated scientific ages (*thick edges* in the figure). From Chen and Redner (2010)

7.4 Modeling Citation Networks

7.4.1 Dynamical Models

Networks of citations between papers are growing systems with complex topological features: the rate at which new papers are added (published) to the network is almost exponential, while the number of references per paper (out-degree) and the number of citations received (in-degree) are broadly distributed. One of the most surprising features of the growth of citation networks, discovered already by de Solla Price (Price 1976), is related to the mechanism ruling the assignment of citations: the probability that a paper gets cited is proportional to the number of citations it already has received. This mechanism is the so-called "cumulative advantage," based on which the "rich get richer," already developed by Yule (1925) and Simon (1957) in different contexts. The criterion, now widely referred to as "preferential attachment," was recently made popular by Barabási and Albert (1999), who proposed it as a general criterion for the emergence of heterogeneous connectivity patterns in networks generated for the description of systems belonging to different scientific domains.

The model by Price (1976) anticipated the modern models of network growth. It is very simple: one node (paper) is introduced (published) at each stage of the growth carrying new connections (citations). The average number of citations (mean degree) is m. The rate at which older nodes receive incoming connections is assumed to be linearly proportional to the number of arcs already incident on them and can be simply indicated by $\Pi\left(k^{in}\right) \sim \left(1+k^{in}\right)$. When a sufficiently large number of papers has been published, the probability that an article has received k^{in} citations becomes stable and, in the limit of large in-degrees, equals

$$P\left(k^{in}\right) \sim \left(k^{in}\right)^{-2-1/m}, \tag{7.2}$$

which means a power law (or "scale free") distribution with exponent $2+1/m$. The exponent of the distribution γ depends on the mean degree m and can therefore be tuned rather arbitrarily.

The Barabási–Albert model (Barabási and Albert 1999), in its standard version, considers the total degree, not the in-degree, and yields a power law degree distribution with $\gamma = 3$. Its extension to the directed case is essentially equivalent to the Price model: the attachment rate is $\Pi\left(k^{in}\right) \sim \left(A+k^{in}\right)$, where $A > 0$ is a parameter that can be tuned (Krapivsky et al. 2000; Dorogovtsev et al. 2000b). In this case, one has $\gamma = 2 + A/m$, where m indicates the number of new citations introduced by each new paper. The exponent $\gamma = 3$ is recovered by setting $A = m$. The preferential attachment model and its subsequent generalizations not only can predict that the tail of the probability distribution for citations follows a power law, but also that the tail will be predominantly composed of the earliest published papers. This effect, supported by empirical evidence and nicely denominated as "first-mover advantage" (Newman 2009), reveals that in order to be well cited, it is often more convenient to write one of the first papers in a particular topic than the best article in that area.

However, the predominant weakness of the preferential attachment model and its variants is the sensitivity to the assumption that the probability of being cited is simply proportional to the number of citations previously collected. One might consider the general ansatz $\Pi\left(k^{in}\right) \sim \left(k^{in}\right)^{\beta}$ for the attachment probability, with a generic β. The scale-free behavior of $P\left(k^{in}\right)$ is observed only for $\beta = 1$: for $\beta < 1$, the distribution of citations turns out to be a stretched exponential, and for $\beta > 1$, a condensation of citations is observed and few papers are cited by nearly all other articles (Krapivsky et al. 2000; Dorogovtsev et al. 2000b).

The preferential attachment hypothesis has undergone empirical validation. Jeong et al. (2003) considered papers published in *Physical Review Letters* in 1988 and all citing articles published later. They divided the time axis into several bins and tested whether the number of citations received up to a certain time was influencing the number of citations received later (Fig. 7.6). They found that papers are cited with a probability that is nearly a linear function of the number of already-received citations, $\Pi\left(k^{in}\right) \sim \left(k^{in}\right)$. A similar result was also observed

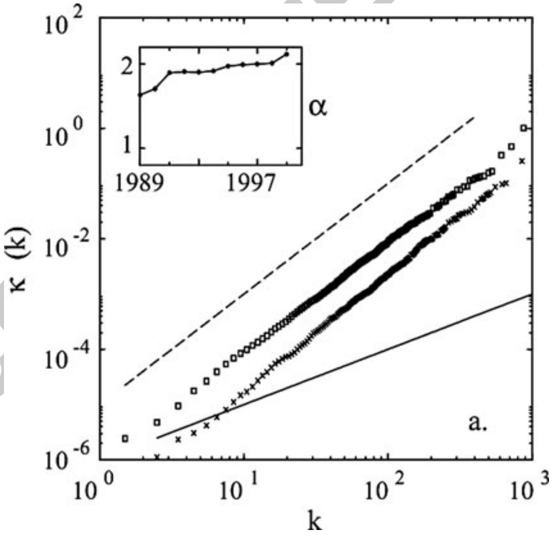


Fig. 7.6 Empirical verification of the validity of the preferential attachment mechanism for citation networks. The cumulative attachment probability $\kappa(k^{in}) = \int_0^{k^{in}} \Pi\left(k^{in}\right)$ should scale as $(k^{in})^{\alpha+1}$ if the original attachment probability $\Pi\left(k^{in}\right)$ scales like $(k^{in})^{\alpha}$ (and vice versa). The cumulative probability $\kappa(k^{in})$ is more suitable than $\Pi\left(k^{in}\right)$ for the empirical analysis because the integral considerably reduces the fluctuations. The two empirical curves correspond to citations received in 1991 and 1995, respectively, by papers published in *Physical Review Letters* in 1988. In both cases $\kappa(k^{in}) \sim (k^{in})^2$, so $\alpha \sim 1$, as in linear preferential attachment. From Jeong et al. (2003)

by Redner (2005) by analyzing the whole dataset of publications in journals of the American Physical Society. Therefore, a linear attachment probability seems to be a typical characteristic of the evolution of citation networks.

An important effect not included in the preferential attachment mechanism is the fact that the probability of receiving citations is time dependent. In the Price model, papers continue to acquire citations independently of their age, while it is reasonable to think and has been empirically observed (Hajra and Sen 2004a,b, 2005; Wang et al. 2008) that the probability for an article to be cited decreases as the age of the same article increases. Some recent papers about growing network models include the aging of nodes as a key feature (Hajra and Sen 2005; Wang et al. 2008; Dorogovtsev and Mendes 2000a, 2001; Zhu et al. 2003). The probability that a paper receives a citation from a new article can be written as $\Pi\left(k^{in}, t\right)$, with explicit dependence not only on the number of citations k^{in} already received but also on the publication time t. For simplicity, the two effects are generally considered independent of each other, and the rate at which papers receive citations becomes separable $\Pi\left(k^{in}, t\right) \sim K\left(k^{in}\right) \cdot f(t)$. Various models have been studied by assuming different functional forms for $K\left(k^{in}\right)$ and $f(t)$. In Dorogovtsev and Mendes (2000a) for example, $K\left(k^{in}\right) = k^{in}$ and $f(t) = t^{\alpha}$. When $\alpha < 0$, the aging effect competes with the preferential attachment mechanism, while for $\alpha > 0$, older nodes are more favored and the age dependence enhances the "rich get richer" effect. The distribution of the number of citations received continues to be a power law for values of $\alpha \geq -1$. In Zhu et al. (2003), $K\left(k^{in}\right) = k^{in}$ and $f(t) = e^{\alpha t}$. The model produces power law distributions for the citations only for $\alpha \leq 0$. A more complicated situation is studied in Dorogovtsev and Mendes (2001), where $K\left(k^{in}\right) = \left(k^{in}\right)^{\beta}$ and $f(t) = t^{\alpha}$. The limiting distributions for the number of citations are studied in the α–β plane: scale-free distributions arise only along the line $\beta = 1$; for $\beta > 1$, condensation phenomena happen and a few nodes acquire almost all the citations; for $\beta < 1$ and $\alpha \leq -1$, the distribution is a stretched exponential.

7.4.2 Static Models

Citation networks are directed and, in good approximation, acyclic graphs. The simultaneous presence of directions and a lack of cycles requires the introduction of specific models able to capture the topological properties of citation networks.

These two ingredients are the basis of the theoretical formulation developed by Karrer and Newman (2009a,b), where the statistical properties of static acyclic and directed graphs are analyzed in detail. Suppose we have a network composed of N articles (nodes) and that the indices of the nodes are chronologically sorted according to their publication date: $j < i$ means that paper j has been published before paper i. Imagine that both the in- and out-degree sequences of the network are given. This means that the number k_i^{in} of papers citing the ith article as well as the number k_i^{out} of publications cited by paper i are completely specified. The study by Karrer and Newman focuses on the statistical properties of the ensemble of networks that can be constructed by preserving the constraint that all incoming and outgoing stubs are paired, with the restriction that only connections of the type $i \rightarrow j$ with $i > j$ are allowed. This static model is very similar to the one represented by the popular configurational model (Molloy and Reed 1998). A natural variable, fundamental for the analytical treatment of the model by Karrer and Newman, is

$$\lambda_i = \sum_{j=1}^{i-1} k_j^{in} - \sum_{j=1}^{i} k_j^{out}, \tag{7.3}$$

which represents the number of incoming stubs "below" node i still available for connections with outgoing stubs exiting from vertices "above" i. In other words, λ_i counts the number of edges that flow "around" the node i. A necessary and sufficient condition for the construction of the model, assuming that all incoming and outgoing stubs are paired in a way that preserves ordering, is that $\lambda_i \geq 0$, $\forall\ 1 < i < N$, while $\lambda_1 = \lambda_N = 0$ arise as the natural boundary conditions of the problem. The expected number of connections between nodes i and j can be estimated to be

$$P_{ij} = k_i^{in} k_j^{out} \frac{\prod_{l=i+1}^{j-1} \lambda_l}{\prod_{l=i+1}^{j} \left(\lambda_l + k_l^{out}\right)}, \tag{7.4}$$

for any pair $i < j$, while $P_{ij} = 0$ otherwise. When the network size grows, P_{ij} becomes small and can be considered equal to the probability of observing a citation from j to i.

The model by Karrer and Newman can reproduce some non-trivial properties of real citation networks (Fig. 7.7) and may provide a useful *null model* for testing topological properties of real citation networks including correlations and modular structures. The model by Karrer and Newman is not able to reproduce a very important topological feature of citation networks, represented by a high occurrence of local triangular structures (Milo et al. 2002). A simple modification of the rules governing the way in which connections are introduced in the network is able to correct this problem. The model by Wu and Holme (2009) is very similar in spirit to the one by Karrer and Newman, but adds two new fundamental ingredients. First, the probability that paper i cites paper j is no longer dependent only on topological and time constraints, but is inversely proportional to the age difference between the two papers (aging effect). Second, once the connection between i and j has been established, there is a finite probability that i copies citations from j and therefore creates triangles. The simultaneous presence of these very intuitive and natural ingredients makes the model more representative of real citation networks.

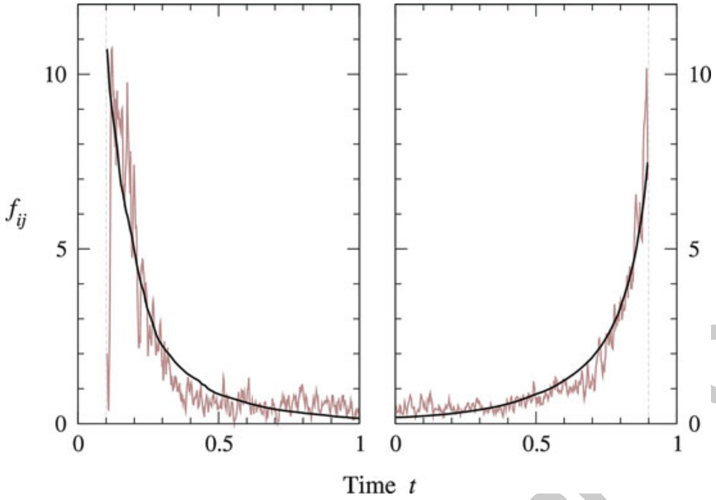


Fig. 7.7 Comparison of the static model by Karrer and Newman with empirical data. One focuses on the function f_{ij}, which is proportional to the connection probability of vertices i and j. The dataset is a citation network of papers on high-energy theory posted on the online eprint archive ArXiv[16] between 1992 and 2003. Papers are ordered from the oldest to the newest. The time of paper i is i/N, and N is the total number of papers. The *left panel* deals with citations from papers at time $t > 0.1$, the *right panel* with citations from papers at time $t < 0.9$. From Karrer and Newman (2009a)

7.5 Dynamics on Citation Networks

Traditional citation metrics, which are used to assess the relevance or popularity of papers, scientists, and journals, rely only on local properties of citation networks. These measures are based on the number of incoming connections of a paper. Simple citation counts quantify the popularity or success of a paper. The number of citations acquired by papers are then transferred to journals and scientists for judgments on their quality. The relevance of journals is quantified by the number of citations received by articles published in them, while the scientific reputation of scientists is measured by the number of citations their articles have received. Even very popular bibliometric indicators, such as the impact factor (Garfield 1955) or the h-index (Hirsch 2005), are based only on purely local properties of citation networks.

Since complete citation networks are currently at our disposal, we can use their entire structure for the formulation of more sophisticated bibliometric measures. Citation networks basically contain information about the dissemination of notions and theories in science, so they may therefore be used as the underlying structures of diffusion processes, where the diffusing particles are nothing more than scientific ideas. The process can be formulated in a straightforward manner where units of scientific credit, carried by papers, diffuse over the network. The generic paper i distributes its credit homogeneously among its k_i^{out} outgoing connections, corresponding to its cited articles. The cited articles will increment their scientific credit by a factor proportional to $1/k_i^{out}$, but then each of these papers will redistribute its total credit to all cited articles, and so on. The entire diffusion process can be mathematically described at the local level, by the following equation

$$P_i = \frac{q}{N} + (1-q) \sum_j \frac{a_{ji}}{k_j^{out}} P_j , \tag{7.5}$$

valid for all $i = 1, \ldots, N$, with N total number of papers in the network. P_i stands for the fraction of scientific credit present on the node i. The increment of P_i is due to two different contributions, one having weight q and the other $1-q$. The first contribution is global and does not depend on the network structure; each paper receives an equal fraction, $1/N$, of scientific credit from the system. Even if by an infinitesimal amount, each and every paper contributes to the scientific advance of a field and is entitled to an infinitesimal $(1/N)$ scientific credit. The second contribution is represented by the flux of credit arriving from the citing papers (the matrix element a_{ji} is one only if j is citing i, while it is zero otherwise). Under general conditions, there is a unique solution for (7.5). The solution can be obtained by starting from suitable initial conditions and then iterating the set of the N equations until each P_i converges to a stable value within an a priori fixed precision. The solution depends on the model parameter q, ranging in the interval [0, 1] and generally called the "damping" or "teleportation" factor. The quantity P_i can be interpreted as a popularity score to be attributed to the paper i in the network.

The method described so far is the same as PageRank (Brin and Page 1998), currently used by the Web search engine *Google* in order to quantify the popularity of web pages. The score assigned to papers is on average linearly proportional to the number of citations received (Fortunato et al. 2008), but large deviations from the average are possible. Papers with high citation counts may have low ranks, while articles with few citations may have high ones. Since the entirety of information of the citation network is used, it is not important merely to be cited many times; the source of citations becomes much more relevant. A single citation from a paper with a high score can be much more important than many citations received by papers with low scores.

In the following, we list the main applications of PageRank's style algorithms to citation networks. It should be stressed that there are not fundamental differences between the various methods since all of them are based on a diffusion process, i.e., (7.5). The differences regard mainly the type of elements ranked according to the diffusion algorithm and, therefore, the application of PageRank algorithm to different types of citation networks.

7.5.1 Ranking of Papers

Chen et al. (2007) applied the former idea to the citation network between papers published in journals from the collection of *Physical Review* from 1893 to 2003.

Using the score obtained from (7.5) with damping factor $q = 0.5$, they were able to identify "gems" among physics papers, not visible from the mere citation count.

A more sophisticated method, based on the same bibliographic dataset, led Walker et al. (2007) to formulate the so-called "CiteRank" score.[17] In CiteRank, the approach based on (7.5) is enriched. Credits still diffuse among the nodes of the citation network, but the diffusion probability has an exponential suppression in time, which prevents credits originating in recent papers from diffusing to much older papers.

7.5.2 *Ranking of Journals*

The diffusion approach is also the key feature of the so-called "Eigenfactor" score,[18] based on which the influence of scientific journals is assessed. In the original formulation of Eigenfactor (Bergstrom 2007; Bergstrom et al. 2008), the authors considered the dataset of *Journal Citation Reports* and constructed the network of citations between all journals indexed by Thompson Scientific. The Eigenfactor score of a journal is an estimation of the percentage of time that library users spend on that journal. The diffusion process of (7.5) here is interpreted as a simple model of bibliographic search, in which readers follow chains of citations as they move from journal to journal. The Eigenfactor score has started to be widely accepted in the scientific community and is at the moment one of the most concrete alternatives to the impact factor.

Analogous to the Eigenfactor, the Science Journal Ranking (SJR) indicator (González-Pereira et al. 2009) represents a bibliometric measure, based on a diffusion algorithm, for the quantification of the prestige of scientific journals. The main difference with respect to the Eigenfactor is the source of bibliographic data, provided in this case by the database Scopus of Elsevier. The SJR indicator is part of the SCImago project, which uses similar bibliometric measures also for the scientific ranking of countries.[19]

7.5.3 *Ranking of Scientists*

A recent approach, still based on a diffusion process, is the Science Author Rank Algorithm (SARA) proposed by Radicchi et al. (2009). The focus of SARA is to assess the impact of scientists and monitor their evolution over time. Given a weighted network of citations between scientists, the score assigned to each author i

---

[17] Citerank, URL: http://www.cmth.bnl.gov/∼maslov/citerank/index.php.

[18] URL: http://www.eigenfactor.org.

[19] SCImago, URL: http://www.scimagojr.com.

is calculated by iterating the set of equations

$$P_i = (1-q)\sum_j \frac{w_{ji}}{s_j^{out}} P_j + qz_i + (1-q)\sum_j P_j\, \delta\left(s_j^{out}\right). \tag{7.6}$$

Equation (7.6) represents the analogue of (7.5) in the case of weighted networks. The first term of the r.h.s. represents the diffusion contribution in the weighted network. Here, the unweighted matrix element a_{ji} is replaced by its weighted version w_{ji}, and the number of outgoing connections k_j^{out} is replaced by the out-strength $s_j^{out} = \sum_i w_{ji}^{out}$. Instead of being redistributed homogeneously, the scientific credits here are drawn back to scientists with probability (z_i) proportional to their scientific productivity (i.e., number of papers published). The last term of the r.h.s. corrects the boundary effects by redistributing the credits of scientists with no outgoing connections to the rest of the network [$\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ for any $x \neq 0$].

The evolution of SARA scores can be monitored by constructing time-dependent networks, where only papers published in a certain time range are used for the construction of the weighted network of citations between scientists. In order to suppress time dependencies in the bare numbers P_i, the rank is constructed on the relative quantity $R_i = 1/N \sum_j \theta\left(P_j - P_i\right)$, which quantifies the probability of finding another author with a SARA rank higher than P_i [$\theta(x) = 1$ if $x > 0$ and $\theta(x) = 0$ for $x < 0$].

Radicchi et al. consider the practical application of their ranking procedure in the case of papers published in journals of the American Physical Society between 1893 and 2006.[20] The authors quantitatively tested the performances of SARA against those of more traditional ranking schemes, such as citation counts. The test was performed on the list of winners of the major prizes in Physics: Nobel Prize, Wolf Prize, Boltzmann Medal, Planck Medal, and the Dirac Medal. By comparing the ranks of these famous scientists based on their SARA scores with those obtained with other measures, the SARA score appears to have a higher predictive value than standard bibliometric indicators like (e.g., citation counts).

7.6 Summary

The massive citation datasets currently available and the need to assess quantitatively the scientific performance of scholars, departments, and universities make the study of citation networks more pressing and germane than ever (even though the assumption that citations represent a proxy for the quantification of scientific relevance may be questionable (Adler et al. 2009)). Citations may occur for many different reasons (Bornmann and Daniel 2008), and papers may stop to receive

---

[20] Phys Author Rank Algorithm, URL: http://physauthorsrank.org.

citations because they become obsolete or textbook material. These factors clearly play important roles and impact the structure and dynamics of citation networks. Since the seminal paper by Price (1965), this field has witnessed an explosive growth, especially in the last decade, and a number of features are now quite well understood.

The distribution of the number of citations received by a paper is broad, although there is still much debate about the actual shape of the distribution. In fact, the shape of the distribution is probably an ill-defined issue, as the distribution may depend on the specific dataset at hand, and the way data are put together. For instance, distributions may be different if one considers papers of the same age or spanning a long period of time, in which case productivity trends may play a role in the final distribution of citations. Furthermore, one has to distinguish the citation habits of different scientific communities. Scholars working on citation networks are now well aware of these issues and important advances are to be expected in the next few years.

The main models for the evolution of citation networks, based on the cumulative advantage rule originally proposed by Price (1976), and cast in a broader perspective by Barabási and Albert (1999), seem to capture the basic features of citation networks. Still, refined models are needed to reproduce real networks in more detail. The attractivity of a paper does not depend only on the number of citations collected by the paper, but also on the age of the paper. Moreover, models based on cumulative advantage usually underestimate the number of (undirected) cycles that one observes in citation networks, as well as the degree correlations between the citing paper and the cited paper. Careful empirical analyses may disclose the origin of such features and how they can be implemented in realistic network models.

Citation networks could also be used to classify papers by topic and subtopic, based on their community structure. The latest developments of community detection in networks may in the near future enable one to analyze even the huge networks that can be constructed with the largest citation databases (e.g., *Web of Science*). One may reveal not only the communities, but also their hierarchical organizations, from the most focused fields to the broadest categories. The resulting classification necessarily will be dynamical, given the rapidly evolving structure of the underlying networks. Processes like the birth, growth, and death of topics may be carefully investigated and modeled.

The sheer number of citations is quite poor as a quantitative indicator of performance. One can do much better by exploiting the full structure of the citation network. Prestige measures based on dynamical processes taking place on citation networks, like PageRank (Brin and Page 1998), are promising alternatives and can still be fast and efficiently computed. In the future, one should consider processes that take into account the specific nature of citation networks (e.g., their approximately acyclic structure and the effect of papers' age).

In general, we expect that the main feature characterizing the future investigations of citation networks will be the time dimension. The analyses of empirical datasets will focus more and more on the evolution of networks, and, consequently, it will be possible to perform comparisons of dynamic network models with data to a level of detail yet unreached.

Key points

1. Statistical laws governing citation distributions; dataset dependence and parametrization.
2. Principle of cumulative advantage, characterization of the network structure.
3. Definition of algorithms for the classification of papers into topics and subtopics based on the community structure of citation networks.
4. Definition of PageRank-like algorithm to achieve system-level ranking measures for papers/authors and topics.
5. Dynamics and time evolution singled out as a crucial feature to achieve understanding and predictive power on knowledge diffusion.

References

Adler R, Ewing J, Taylor P (2009) Citation statistics. Statist Sci 24(1):1–14 (DOI: 10.1214/09-STS285), also available as arXiv preprint arXiv:0910.3529

Albert R, Barabási AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74(1):47–97 (DOI: 10.1103/RevModPhys.74.47), also available as arXiv preprint arXiv:cond-mat/0106096

Anastasiadis AD, de Albuquerque MP, de Albuquerque MP, Mussi, DB (2009) Tsallis q-exponential describes the distribution of scientific citations – a new characterization of the impact. Scientometrics 83:205–218 (DOI: 10.1007/s11192-009-0023-0), also available as arXiv preprint arXiv:cond-mat/0812.4296

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512 (DOI: 10.1126/science.286.5439.509), also available as arXiv preprint arXiv:cond-mat/9910332

Barrat A, Barthélemy M, Vespignani A (2008) Dynamical processes on complex networks. Cambridge University Press, Cambridge (DOI: 10.1017/CBO9780511791383)

Bergstrom C (2007) Eigenfactor: Measuring the value and prestige of scholarly journals. College Res Libraries News 68(5(285)):314–316. http://crln.acrl.org/content/68/5/314.full.pdf+html

Bergstrom CT, West JD, Wiseman MA (2008) The Eigenfactor^{TM} metrics. J Neurosci 28(45):11433–11434 (DOI: 10.1523/JNEUROSCI.0003-08.2008)

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. Phys Rep 424(4–5):176–308 (DOI: 10.1016/j.physrep.2005.10.00)

Bommarito II MJ, Katz DM (2009) Properties of the United States code citation network. ArXiv preprint arXiv:0911.1751

Bommarito II MJ, Katz DM, Zelner JL (2010a) On the stability of community detection algorithms on longitudinal citation data. Procedia Soc Behav Sci 4:26–37 (DOI: 10.1016/j.sbspro.2010.07.480), also available as arXiv preprint arXiv:0908.0449. The article is part of: Hirschi C, Ingold K, Serdült U (guest eds) Applications of Social Network Analysis: 6th Conference on Applications of Social Network Analysis. Procedia Soc Behav Sci 4:1–102.

Bommarito II MJ, Katz DM, Zelner JL, Fowler JH (2010b) Distance measures for dynamic citation networks. Physica A 389(19):4201–4208 (DOI: 10.1016/j.physa.2010.06.003), also available as arXiv preprint arXiv:0909.1819

Bornmann L, Daniel HD (2008) What do citation counts measure? A review of studies on citing behavior. J Doc 64(1):45–80 (DOI: 10.1108/00220410810844150)

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 (DOI: 10.1016/S0169-7552(98)00110-X) (The article is part of: Proceedings of the seventh international World Wide Web conference. Computer Networks and ISDN Systems 30(1–7):1–778)

Caldarelli G (2007) Scale-free networks: Complex webs in nature and technology. Oxford Finance Series. Oxford University Press, Oxford

Castellano C, Radicchi F (2009) On the fairness of using relative indicators for comparing citation performance in different disciplines. Archivum Immunologiae et Therapiae Experimentalis 57(2):85–90 (DOI: 10.1007/s00005-009-0014-0)

Chen P, Xie H, Maslov S, Redner S (2007) Finding scientific gems with Google. J Informetr 1(1):8–15 (DOI: 10.1016/j.joi.2006.06.001), also available as arXiv preprint arXiv:physics/0604130

Chen P, Redner S (2010) Community structure of the Physical Review citation network. J Informetr 4(3):278–290 (DOI: 10.1016/j.joi.2010.01.001), also available as arXiv preprint arXiv:0911.0694

Dorogovtsev SN, Mendes JFF (2000a) Evolution of networks with aging of sites. Phys Rev E 62(2):1842–1845 (DOI: 10.1103/PhysRevE.62.1842), also available as arXiv preprint arXiv:cond-mat/0001419

Dorogovtsev SN, Mendes JFF, Samukhin AN (2000b) Structure of growing networks with preferential linking. Phys Rev Lett 85(21):4633–4636 (DOI: 10.1103/PhysRevLett.85.4633), also available as arXiv preprint arXiv:cond-mat/0004434

Dorogovtsev SN, Mendes JFF (2001) Scaling properties of scale-free evolving networks: Continuous approach. Phys Rev E 63(5):056125, 19 pp (DOI: 10.1103/PhysRevE.63.056125), also available as arXiv preprint arXiv:cond-mat/0012009

Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. Adv Phys 51(4):1079–1187 (DOI: 10.1080/00018730110112519), also available as arXiv preprint arXiv:cond-mat/0106144

Fortunato S, Boguña M, Flammini A, Menczer F (2008) Approximating PageRank from in-degree. In: Aiello W, Broder A, Janssen J, Milios E (eds) Algorithms and models for the web-graph: Fourth international workshop, WAW 2006, Banff, Canada, November 30–December 1, 2006: Revised papers. Lecture Notes in Computer Science, vol 4936. Springer, Berlin, pp 59–71 (DOI: 10.1007/978-3-540-78808-9_6), also available as arXiv preprint arXiv:cs/0511016

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174 (DOI: 10.1016/j.physrep.2009.11.002), also available as arXiv preprint arXiv:0906.0612

Garfield E (1955) Citation indexes to science: A new dimension in documentation through association of ideas. Science 122(3159):108–111 (DOI: 10.1126/science.122.3159.108). Reprinted in: Garfield E (1984) Essays of an information scientist: Volume 6. ISI Press, Philadelphia, PA, pp. 468–471, also available online at the URL: http://www.garfield.library.upenn.edu/essays/v6p468y1983.pdf

González-Pereira B, Guerrero-Bote VP, Moya-Anegón F (2009) The SJR indicator: A new indicator of journals' scientific prestige. ArXiv preprint arXiv:0912.4141

Hajra KB, Sen P (2004a) Phase transitions in an aging network. Phys Rev E 70(5):056103, 4 pp (DOI: 10.1103/PhysRevE.70.056103), also available as arXiv preprint arXiv:cond-mat/0406332

Hajra KB, Sen P (2004b) Aging in citation networks. Physica A 346(1–2):44–48 (DOI: 10.1016/j.physa.2004.08.048), also available as arXiv preprint arXiv:cond-mat/0409017. The article is part of: Manna SS, Sen P (guest eds) Statphys – Kolkata V: Proceedings of the International Conference on Statistical Physics: 'Complex networks: Structure, function and processes' Kolkata, India, 27 June–01 July 2004. Physica A 346(1–2)

Hajra KB, Sen P (2005) Modeling aging characteristics in citation networks. Physica A 368(2):575–582 (DOI: 10.1016/j.physa.2005.12.044), also available as arXiv preprint arXiv:physics/0508035

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569–16572 (DOI: 10.1073/pnas.0507655102, stable JSTOR URL: http://www.jstor.org/stable/4152261), also available as arXiv preprint arXiv:physics/0508025

Jain AK, Murty MN, Flynn PJ (199) Data clustering: a review. ACM Comput Surveys 31(3):264–323 (DOI: 10.1145/331499.331504)

Jeong H, Neda Z, Barabási AL (2003) Measuring preferential attachment in evolving networks. Europhys Lett 61(4):567–572 (DOI: 10.1209/epl/i2003-00166-9), also available as arXiv preprint arXiv:cond-mat/0104131

Karrer, B, Newman MEJ (2009a) Random acyclic networks. Phys Rev Lett 102(12):128701, 4 pp (DOI: 10.1103/PhysRevLett.102.128701), also available as arXiv preprint arXiv:0902.4013

Karrer B, Newman MEJ (2009b) Random graph models for directed acyclic networks. Phys Rev E 80(4):046110, 12 pp (DOI: 10.1103/PhysRevE.80.046110), also available as arXiv preprint arXiv:0907.4346

Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. Phys Rev Lett 85(21):4629–4632 (DOI: 10.1103/PhysRevLett.85.4629), also available as arXiv preprint arXiv:cond-mat/0005139

Kryssanov VV, Kuleshov EL, Rinaldo FJ, Ogawa H (2007) We cite as we communicate: A communication model for the citation process. ArXiv preprint arXiv:cs/0703115

Laherrére J, Sornette D (1998) Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales. Eur Phys J B 2(4):525–539 (DOI: 10.1007/s100510050276), also available as arXiv preprint arXiv:cond-mat/9801293

Lehmann S, Lautrup B, Jackson AD (2003) Citation networks in high energy physics. Phys Rev E 68(2):026113, 8 pp (DOI: 10.1103/PhysRevE.68.026113), also available as arXiv preprint arXiv:physics/0211010

Leicht E A, Clarkson G, Shedden K, Newman MEJ (2007) Large-scale structure of time evolving citation networks. Eur Phys J B 59(1):75–83 (DOI: 10.1140/epjb/e2007-00271-7), also available as arXiv preprint arXiv:0706.0015

Leicht EA, Newman MEJ (2008) Community structure in directed networks. Phys Rev Lett 100(11):118703, 4 pp (DOI: 10.1103/PhysRevLett.100.118703), also available as arXiv preprint arXiv:0709.4500

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–323

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: Simple building blocks of complex networks. Science 298(5594):824–827 (DOI: 10.1126/science.298.5594.824)

Molloy M, Reed B (1998) The size of the largest component of a random graph on a fixed degree sequence. Combinatorics Probability Comput 7(3):295–306 (DOI: 10.1017/S0963548398003526)

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256 (DOI: 10.1137/S003614450342480), also available as arXiv preprint arXiv:cond-mat/0303516

Newman MEJ (2009) The first-mover advantage in scientific publication. Europhys Lett 86(6):68001 (DOI: 10.1209/0295-5075/86/68001), also available as arXiv preprint arXiv:0809.0522

Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the Internet: A statistical physics approach. Cambridge University Press, Cambridge (DOI: 10.1017/CBO9780511610905)

Price DJ de Solla (1965) Networks of scientific papers. Science 149(3683):510–515 (DOI: 10.1126/science.149.3683.510). Reprinted in: Price DJ de Solla (1986) Little science, big science . . . and beyond. Columbia University Press, New York, NY, pp 103–118

Price DJ de Solla (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306 (DOI: 10.1002/asi.4630270505), also available online at the URL: http://www.asis.org/Publications/JASIS/Best_Jasist/1976pricejasistarticle.pdf

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Towards an objective measure of scientific impact. Proc Natl Acad Sci USA 105(45):17268–17272 (DOI: 10.1073/pnas.0806977105, stable JSTOR URL: http://www.jstor.org/stable/25465266), also available as arXiv preprint arXiv:0806.0974

Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80(5):056103, 10 pp (DOI: 10.1103/PhysRevE.80.056103), also available as arXiv preprint arXiv:0907.1050

Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Eur Phys J B 4(2):131–134 (DOI: 10.1007/s100510050359), also available as arXiv preprint arXiv:cond-mat/9804163

Redner S (2005) Citation statistics from 110 years of Physical Review. Phys Today 58(6):49–54 (DOI: 10.1063/1.1996475), also available as arXiv preprint arXiv:physics/0506056

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118–1123 (DOI: 10.1073/pnas.0706851105, stable JSTOR URL: http://www.jstor.org/stable/25451237), also available as arXiv preprint arXiv:0707.0609

Scott J (200) Social network analysis: A handbook, 2nd edn. Sage Publications, London

Seglen PO (1999) The skewness of science. J Am Soc Inf Sci 43(9):628–638 (DOI: 10.1002/(SICI)1097-4571(199210)43:9¡628::AID-ASI5¿3.0.CO;2-0)

Shockley W (1957) On the statistics of individual variations of productivity in research laboratories. Proc IRE 45(3):279–290 (DOI: 10.1109/JRPROC.1957.278364)

Simkin MV, Roychowdhury VP (2005) Copied citations create renowned papers? Annals Improbable Res 11(1):24–27, available online at the URL: http://improbable.com/airchives/paperair/volume11/v11i1/v11i1.html, also available as arXiv preprint arXiv:cond-mat/0305150

Simon HA (1957) Models of man: Social and rational: Mathematical essays on rational human behaviour in a social setting. Wiley, New York, NY; Chapman & Hall, London

Stringer MJ, Sales-Pardo M, Amaral LAN (2008) Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE 3(2):e1683, 8 pp (DOI: 10.1371/journal.pone.0001683)

Stringer MJ, Sales-Pardo M, Amaral LAN (2010) Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. J Am Soc Inf Sci Techol 61(7):1377–1385 (DOI: 10.1002/asi.21335)

Tsallis C, de Albuquerque, MP (2000) Are citations of scientific papers a case of nonextensivity? Eur Phys J B 13(4):777–780 (DOI: 10.1007/s100510050097), also available as arXiv preprint arXiv:cond-mat/9903433

van Raan AFJ (2001a) Two-step competition process leads to quasi power-law income distributions: Application to scientific publications and citation distributions. Physica A 298(3–4):530–536 (DOI: 10.1016/S0378-4371(01)00254-0)

van Raan AFJ (2001b) Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions. Scientometrics 51(1):347–357 (DOI: 10.1023/A:1010501820393)

Vazquez A (2001) Statistics of citation networks. ArXiv preprint arXiv:cond-mat/0105031

Walker D, Xie H, Yan KK, Maslov S (2007) Ranking scientific publications using a model of network traffic. J Stat Mech 2007(6):P06010, 10 pp (DOI: 10.1088/1742-5468/2007/06/P06010), also available as arXiv preprint arXiv:physics/0612122

Wallace ML, Lariviére V, Gingras Y (2009) Modeling a century of citation distributions. J Informetr 3(4):296–303 (DOI: 10.1016/j.joi.2009.03.010), also available as arXiv preprint arXiv:0810.1426

Wang M, Yu G, Yu D (2008) Measuring the preferential linear attachment mechanism in citation networks. Physica A 387(18): 4692–4698 (DOI: 10.1016/j.physa.2008.03.017)

Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Structural Analysis in the Social Sciences, vol 8. Cambridge University Press, Cambridge

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442 (DOI: 10.1038/30918)

Wu ZX, Holme P (2009) Modeling scientific-citation patterns and other triangle-rich acyclic networks. Phys Rev E 80(3):037101, 4 pp (DOI: 10.1103/PhysRevE.80.037101), also available as arXiv preprint arXiv:0908.2615

Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Wills, F.R.S.. Proc R Soc Lond B Biol Sci 213(403):21–87 (DOI: 10.1098/rstb.1925.0002, stable JSTOR URL: http://www.jstor.org/stable/92117)

Zhu H, Wang X, Zhu JY (2003) Effect of aging on network structure. Phys Rev E 68(5):056121, 9 pp (DOI: 10.1103/PhysRevE.68.056121), also available as arXiv preprint arXiv:cond-mat/0304609

AUTHOR QUERIES

AQ1. Non-bibliographic references have been moved as text footnote. Please check.

AQ2. Please check whether the inserted citation for Table 7.1 is appropriate.

Part IV
Outlook

Chapter 8
Science Policy and the Challenges for Modeling Science

Peter van den Besselaar, Katy Börner, and Andrea Scharnhorst

8.1 Challenges and Opportunities

This book seeks to advance the modeling science to improve our collective understanding of the functioning of science systems and of the dynamics of science. It also attempts to make the modeling of science relevant from the perspective of AQ1 societal use – an issue that is increasingly important in scientific research.

In the last decade, we have witnessed a renewed interest among science policy-makers in the science of science and of science policy (Executive Office 2008). In several countries, new programs and institutes have been established to study the dynamics of science with an explicit application orientation.[1] The results of these research activities are expected to inform science policy-makers in different positions: within national government, within research councils and other agencies

---

[1] For example, the center for Science System Assessment in the Netherlands, the Institute for Research Information and Quality Assurance (IFQ) in Germany, the NSF Science of Science and Innovation Policy program in the US, the former Prime Network of Excellence in the EU.

P. van den Besselaar (✉)
Department of Organization Science & Network Institute, VU University Amsterdam,
The Netherlands
e-mail: p.a.a.vanden.besselaar@vu.nl

K. Börner
SLIS, Indiana University, Bloomington, IN
e-mail: katy@indiana.edu

A. Scharnhorst
Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences,
Anna van Saksenlaan 10, 2593 HT Den Haag, The Netherlands
e-mail: andrea.scharnhorst@dans.knaw.nl

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4_8,

active in research agenda setting and research funding, and within universities and public research institutes.

What type of knowledge would science policy-makers need, and what is the role of modeling in this context? Three broad classes of questions can be distinguished. Firstly, science policy has a need for the dynamic monitoring and forecasting of scientific developments and technological breakthroughs. They are interested in recognizing promising developments in existing and new research fields early in support of agenda setting and investment decisions. Secondly, there is a need for better understanding the institutional and organizational conditions for a healthy and high-performing research system. How should the research system be organized to realize the heterogeneous goals that come with research? What funding arrangements function effectively under which conditions? How should research evaluation be organized in order to improve performance of the research system, organizations, and researchers? Thirdly, scientific knowledge is increasingly crucial for innovation and societal problem solving. This, together with the rising investments in research, increases the pressure on researchers, research organizations, and research funders to show that their activities do have a societal impact. How should the interaction between knowledge producers and (potential) knowledge users be organized in order to maximize societal impact? What incentives may be implemented to improve these interactions, without destroying the independence and autonomy of science that are crucial for the long-term growth of knowledge? And, what metrics could be developed to measure and show this impact?

These three domains of science policy problems (forecasting the dynamics of science, accelerating research, and improving and measuring the societal impact of research) can be translated into a broad research agenda for the science of science.

To be truly useful for informing science policy, such a research agenda should not only be analytically *divided* into a large set of research questions focusing on specific issues. There is also a strong quest for *synthesis*, for integrating the knowledge obtained about the various different relevant mechanisms. From a policy perspective, one is not primarily interested in the individual mechanisms, or in the relations between small sets of variables, but in the working of the research system as a whole with its many heterogeneous relations between many heterogeneous agents. This asks for mixed-method, multi-level models of the science systems (Börner et al. 2010) that help to understand the relevant processes, dynamics, and complex interactions and their outcomes. Science policy needs a synthetic approach next to analytical approaches to study separate dimensions of science, science dynamics, and the science system.

Such a systems approach to science and science policy studies is becoming possible because of three developments:

1. Firstly, new methodologies of modeling the dynamics of networks of scientific information have been developed. Detailed models of science are becoming available that help to understand the relevant processes, dynamics, and complex interactions and their outcomes.

2. Secondly, testing complex models requires large amounts of high-quality and high-coverage data. Fortunately, new types of digital data are becoming available for studying the structure, organization, and development of science. Among them are survey datasets, and also new bibliographic and other databases, leading to a growing system of "linked open data" and semantic web technologies that enable the integration and use of these data for research (Berners-Lee and Fischetti 1999; Berners-Lee et al. 2001; Heath and Bizer 2011). Many datasets are crowd-sources by thousands using collaborative tools such as CiteULike[2] or Mendeley,[3] but also more generally the WWW and a large variety of existing data sources.
3. Finally, complex models and large-scale data analysis require new methodologies and tools for visualizing and communicating results. Major progress has been made over the last decade (Börner 2010), among others tools for data analysis and visualization available in researcher networking support sites such as VIVO[4] and Collexis,[5] as well as in Scholarometer[6] and author-mapping tools.[7]

8.2 Contributions of this Book

This book provides a review of major methodologies of modeling the dynamics of networks of scientific information, many of which seem to have promising applications in science and science policy studies. The chapters in this book review major models, but not all modeling branches and possible approaches have been covered. Although the team of editors and authors underwent extensive efforts to link the chapters to each other and to use re-occurring elements – such as listings of covered models and their main contributions in the beginning of the chapter and take-away boxes at the end – each chapter comes with its own style and language expressing the different epistemic cultures and traditions in which each specific author feels at home. A variety of knowledge- domain-specific vocabulary and mathematical languages can be found.

This points to an open problem that this first review of major models of science does not manage to solve: the necessity of translation and mutual mapping. The mathematical translation of the different models is as challenging as their conceptual translation and integration. Possible dimensions along which the models presented in the book can be related to each other comprise:

---

[2] URL: http://www.citeulike.org.

[3] URL: http://www.mendeley.com.

[4] URL: http://www.vivoweb.org.

[5] URL: http://www.collexis.com.

[6] URL: http://scholarometer.indiana.edu.

[7] URL: http://www.authormapper.com.

- Units/parts of the science system that a model aims to reproduce.
- Questions that the model aims to answer.
- Mathematical approaches used.
- Visualizations employed to communicate results.
- Insights gained.

The comparison of different models for means of validation and their synergistic combination to increase the quality and coverage of models for capturing the science system require future research.

8.3 Future Work

The various models of science and science dynamics not only use different mathematical approaches to model science, but they also capture different aspects of science and its dynamics. Therefore, integrating models is not only a mathematical task, but is also at the same time an effort to define and combine the different conceptual and theoretical mechanisms specified by individual models.

Most efforts to model aspects of science focus on modeling knowledge spaces and information spaces, and their dynamics – missing are the social and organizational aspects of knowledge production (van den Besselaar 2011). Social behavior of agents in the models is often (but not always) very stylized, and does not represent the richness of aims, interests, strategies, resource distributions, and rules that characterize science. As argued above, from a science policy perspective one is not only interested in modeling and mapping scientific information and the dynamics of science. A second important and still open problem this book only addresses marginally is modeling (i) those social and organizational factors influencing knowledge production and knowledge dynamics, and (ii) the interactions between knowledge growth and knowledge use, including the social characteristics of these links between researchers and research institutions on the one hand, and the users of knowledge on the other.

This leads to a second challenge for future science modeling research. There is a need not only for integrating the existing models that focus on knowledge dynamics and co-author patterns, but also for capturing the different processes in the science system. It is useful to distinguish three dimensions of the science system: *researching*, *codifying*, and *organizing*. *Researching* refers to the everyday practice of doing research, of collaborating and communicating. *Codifying* refers to the output of research, to the publication process where research results are integrated into the existing body of knowledge. Finally, *organizing* refers to all processes for creating the conditions for research at various levels, including but not restricted to science policy.

Most existing models focus on the codifying dimension of science – the communication processes in the formal (journal) literature. Thus, the focus is on the output side, neglecting the underlying processes. Knowledge dynamics is modeled, and generally this only takes into account the underlying social processes of research collaboration – operationalized as co-authoring. The processes of *researching*, however, are only marginally covered (Gilbert 1997; Payette 2011). Here, different kinds of researchers' behavior become relevant, such as *collaboration* in informal and more formal (projects) ways, and informal *communication* in a multitude of forms, such as face-to-face and a variety of social media (research blogs, email lists, etc.). Data about this dimension of researchers' behavior becomes increasingly accessible, as much behavior leaves digital traces in the used social media. This refers to the second of the three developments mentioned in the first section: new types of data are becoming available for the science of science.

Models that aim to capture the research process might help answer questions such as "How to create productive teams?" or "Where do innovations come from?" and not only where they are located in the formal communication spaces of journals and papers. They will make it possible to study the interaction between research communication and collaboration on the one hand and the formal scholarly communication and publication on the other. If successful, this line of modeling might be able to relate performance indicators, such as counts of publications and citations, Crown-indicators and H-indices, to the underlying research process Wallace (2009). And, an improved understanding of research processes may help to develop new indicators, which are not necessarily based on publications and citations Alt-metrics (Mendeley Group).[8]

The next challenge is to include processes of *organizing* research (in a broad sense) in modeling efforts: the different modalities of research funding, agenda setting, research evaluation, and selecting researchers and shaping academic careers. Differences within and between science systems impact the behavior of individual researchers and result in vastly different outcomes that have a strong impact on the research profiles and strengths of different organizations and countries.

Thirdly, future science models should study the interactions between researchers and their organizations on the one hand, and (potential) users of knowledge on the other, in order to better understand the processes of uptake and societal use of scholarly knowledge. They should attempt to capture how knowledge flows through complex networks of researchers and knowledge users, and what attributes, behaviors, incentives, and organizational forms have what effects on these flows.

8.4 Conclusions

In this outlook, we sketched briefly a broad agenda for the future development of models of science, an agenda that combines scholarly and science policy relevance. Traditionally, science models have aimed to answer isolated questions about specific aspects of the science system. In the future there is a need for a more synthetic

---

[8] URL: http://www.mendeley.com/groups/586171/alt-metrics.

approach that integrates different models to capture multiple interacting levels of the science system. The research approach has to be multi-theoretical and multi-level – spanning the individual decision making of researchers to the national science policy decisions – to validate these models using the growing availability of (digital) data about the science system, and use increasingly sophisticated methods and tools for visualizing results.

Last but not least, we hope that the different research streams of science modeling in economics, physics, social science, science of science, and other fields of science will get interlinked not only along the arrow of time, through the historical roots they share, but also in the current time slice in which they are located.

References

Berners-Lee T, Fischetti M (1999) Weaving the web: The original design and ultimate destiny of the World Wide Web by its inventor. Harper, San Francisco, SF

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Scientific Am 284(5):28–37

Börner K (2010) Atlas of science: Visualizing what we know. MIT Press, Cambridge, MA

Börner K, Contractor N, Falk-Krzesinski HJ, Fiore SM, Hall KL, Keyton J, Spring B, Stokols D, Trochim W, Uzzi B (2010) A multi-level systems perspective for the science of team science. Sci Transl Med 2(49):49cm24, 5 pp (DOI: 10.1126/scitranslmed.3001399), also available online at the URL: http://ivl.slis.indiana.edu/km/pub/2010-borner-et-al-multi-level-teamsci.pdf

Executive Office of the President of the United States (2008) The science of science policy: A federal research roadmap: Report on the science of science policy to the Subcommittee on Social, Behavioral and Economic Sciences, Committee on Science, National Science and Technology Council, Office of Science and Technology Policy. Report DOE/SC-106. Executive Office of the President of the United States, Washington, DC, available online at the URL: http://www.whitehouse.gov/files/documents/ostp/NSTC Reports/39924_PDF Proof.pdf

Gilbert N (1997) A simulation of the structure of academic science. Sociol Res Online 2(2), available online at the URL: http://www.socresonline.org.uk/2/2/3.html

Heath T, Bizer C (2011) Linked data: Evolving the web into a global data space. Synthesis Lectures on the Semantic Web: Theory and Technology, vol 1. Morgan & Claypool, San Rafael, CA (DOI: 10.2200/S00334ED1V01Y201102WBE001)

Payette N (2011) Agent-based models of science. In: Scharnhorst A, Börner K, van den Besselaar P
AQ2 (eds) Models of science dynamics. Springer, Berlin, Ch 4, pp xx–xx

van den Besselaar P (2011) Postexcellente wetenschap. Inaugural lecture for the Chair “Organisatie
AQ3 en Dynamiek van de Wetenschap”. VU University, Amsterdam, to be published

Wallace ML, Lariviére V, Gingras Y (2009) Modeling a century of citation distributions. J Informetr 3(4):296–303 (DOI: 10.1016/j.joi.2009.03.010), also available as arXiv preprint arXiv:0810.1426

AUTHOR QUERIES

AQ1. Non-bibliographic references have been moved as text footnote. Please check.

AQ2. Please provide page range for reference Payette (2011).

AQ3. Please update reference van den Besselaar (2011).

Index

A. Scharnhorst et al. (eds.), *Models of Science Dynamics*, Understanding Complex Systems, DOI 10.1007/978-3-642-23068-4, © Springer-Verlag Berlin Heidelberg 2012