Chapter1 in "Models of science dynamics - Encounters between complexity theory and information sciences." Springer book in the Understanding Complex Systems series. *Editors: Andrea Scharnhorst, Katy Börner, Peter van den Besselaar.*

An introduction to modeling science: Basic model types, key definitions, and a general framework for the comparison of process models

Katy Börner[1], Kevin W. Boyack[2], Staša Milojević[1], Steven Morris[3]

[1]SLIS, Indiana University, Bloomington, IN, email: katy@indiana.edu, smilojev@indiana.edu
[2]SciTech Strategies, Albuquerque, NM, email: kboyack@mapofscience.com
[3]Baker Hughes, Houston, Texas, email: Steven.Morris@bakerhughes.com

Abstract

A model is a systematic description of an object or phenomenon that shares important characteristics with its real-world counterpart and supports its detailed investigation. Models are commonly represented as a system of postulates, data, and inferences presented visually, in material form, in mathematical terms, or as a computer simulation. They are often used in the construction of scientific theories. Models of science aim to capture the structure and/or dynamics of science itself. Data on the science system—scholars, papers, patents, grants, jobs, etc. and their complex interdependencies and dynamics—are used to validate them. This chapter provides a general introduction to the modeling of science together with a discussion of different model types, basic definitions, and an overview and comparison of major predictive models of science covered in this book. The appendix provides definitions of common terminology.

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, using ever better methods and instruments, to generate new knowledge, which then appears in tangible forms as journal articles, reports, books, patents, data and software repositories, etc (Whitley, 2000).

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 the work of a social ***network of researchers*** that generate and validate a ***network of knowledge***. This definition is based on the premise that science represents 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 the historical, institutional, political, and research contexts. At the same time, the meanings of the concepts 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, Scott, & Gibbons, 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. Most scientometric studies describe the structure and evolution of science while a few others aim to replicate and predict the structure and dynamics of science.

1.1. Science as a social activity

The relationships between scholars and the structure of 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 a government agency, an industry segment, or a university, 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, or quantitatively as stages of cluster formation (Ben-David & Collins, 1966; Crane, 1969; Crane, 1972; Mullins, 1972; Mullins, 1973).

[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.

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, which exercise significant power over rewards and thus shape the behavior of scholars.

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. Given the importance of textual documents in the practice of science (Callon, Courtial, Turner, & Bauin, 1983; Latour, Salk, & Woolgar, 1986), it is natural to focus on the shared conceptual system of scientific communities as expressed through the terminology used in those documents.

Studies of scientific knowledge are mostly interested in the emergence, growth, and spread (diffusion) of scientific ideas. There are different ways in which one can study scientific knowledge. The most common is to study documents or artifacts produced by scholars. 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, e.g., word co-occurrence analysis. Another approach is to treat references as concept symbols and then perform a whole range of analyses using references as a data source. These analyses can be used to produce maps of science which seek to visually describe the structure of the data (Börner, Chen, & Boyack, 2003). The third approach is to take journals as units of analysis and study their subjects. These analyses are often used for studying interdisciplinarity.

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 desire to model a system at multiple levels.

2.1 Definition and general design of a science model

Models are simplified representations of a system. Models attempt to reduce the world to a fundamental set of elements and laws and on this basis they hope to better understand and predict key aspects of the world. Models capturing the structure and dynamics of scientific endeavor are expected to provide insights into 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) 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, reputation in space analogous to food foraging studied by anthropologists. 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., paper published during spring semester or summer break), etc. on diffusion patterns and dynamics. In addition, there are models that simulate the growth of (coupled) 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 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 or 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 in 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 code that attempts to mathematically describe the structure or dynamics of interest. Subsequently, the model is run and validated by comparing simulated data 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 Figure 1.

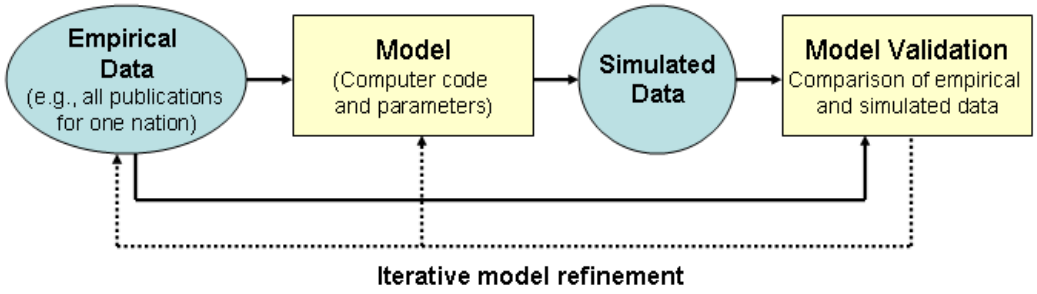

Figure 1. General model design, validation, and refinement process

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.

2.3 Deductive models vs. inductive 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 takes a "bottom up" approach that starts with specific observations and measures, continues with the identification of patterns and regularities, then 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 examination of a set of particulars. It is more open-ended and exploratory, especially at the beginning.

2.4 Deterministic models vs. stochastic models

Deterministic models describe the behavior of an object or phenomenon whose behavior is 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.

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 are ***process models***, which aim to capture the mechanisms and temporal dynamics by which real-world networks are created (Newman & A., 2007; Zhang, Qiu, & Ivanova, 2010), with particular emphasis on identification of elementary mechanisms that lead to the emergence of specific network structures and dynamics. The 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. Respective models are interested in explaining 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 are Lotka's law (1926), Bradford's law (1934), Zipf's law (1949). ***Computational models*** describe the structure of dynamics of science using different computational approaches such as agent based modeling, population models (Bettencourt, Kaiser, Kaur, Castillo-Chavez, & Wojick, 2008), cellular automata, or statistical mechanics.

A number of studies that used co-authorship networks to study network dynamics (Barabasi et al., 2002; Barabási & Albert, 1999; Farkas et al., 2002; Nagurney, 1997; Newman, 2001) reveal the existence of small-world and scale-free network topologies (see section 2.4) and preferential attachment (de Solla 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.

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. An example are scale free network models (Barabási & Albert, 1999) or small world network models (Watts & Strogatz, 1998) that generate 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. An example is Goffman's (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 other dynamic models, one models 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 (de Solla Price, 1963). Since then, growth has been largely linear since then 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. Namely, not only that there are more authors on a single paper, but these authors come from different disciplines, different institutions, and different knowledge production sites (e.g. university and industry). In addition, there is a wide geographic distribution of coauthors as well. This is the result of 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, Maru, & Goldstone., 2004) as well as work that examines the interplay of existing network structures and resulting scholarly dynamics that in turn affect the growth of scholarly networks.

2.7 Multi-level and multi-perspective models

It is often desirable to model a system at multiple levels using different vantage points, see Figure 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.

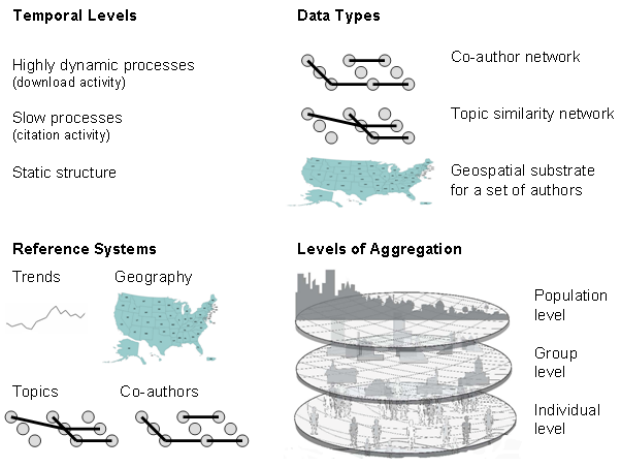

Figure 2. Temporal levels (top left), data types (top right), reference systems (lower left), and levels of aggregation (lower right, Adopted from Vespignani).

2.8 Exemplification using predictive workflows

As mentioned in section 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, make up documents, make up line of research; papers appear in journals, journals group into disciplines or subject categories, 13 major fields, or all of science
- ***with whom (network analysis):*** person is part of research team, part of research community/invisible college; person works at 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). Here 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 decisions makers (from team leaders to university officials to agency heads) that can potentially be informed by science models. These include:

- How do changing resources change 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 existing and new people and resources would be key enablers?

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 the 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, Sharma, & Karisiddappa, 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.

3. Basic Conceptualization and Science Modeling Terminology

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 despite the fact that science models have been designed to answer vastly different questions at many levels of generality. In an attempt to provide a unifying conceptualization (Börner & 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 previously published frameworks by Morris and Rodriguez (Morris & Martens, 2008; Morris & Yen, 2004; Rodriguez, Bollen, & Van de Sompel, 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 section 1.1) will use researchers, teams, invisible colleagues as key *social terms*. Models that simulate science as a knowledge network (see section 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*. 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 that are alive can co-author) and geospatial and topic space (e.g., authors that are geospatially close and work on similar topics are more likely to co-author).

All underlined terms are defined in the appendix and these definitions provide also more information on the concrete interlinkages.

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 grouping of these terms are possible. 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). The three derivative 2-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

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

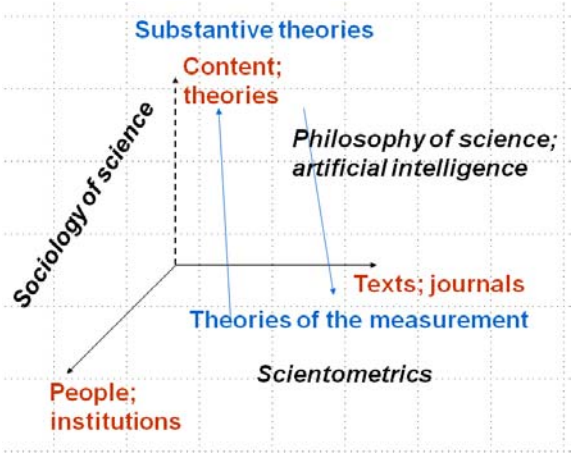

Figure 3. Three-dimensional space by Leydesdorff (1995). The tree axes stand for units of analysis. A phenomenon can be represented as a point in this space with a value on each of the three axes via projection. An example is an institutional rule that would be attributed to an institution, might be represented as text, and has cognitive (substantive) content.

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

A final alternative, network-based approach is given in Figure 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, documents. Conceptual aggregations such as invisible colleges or specialties can be analyzed and mapped. 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.

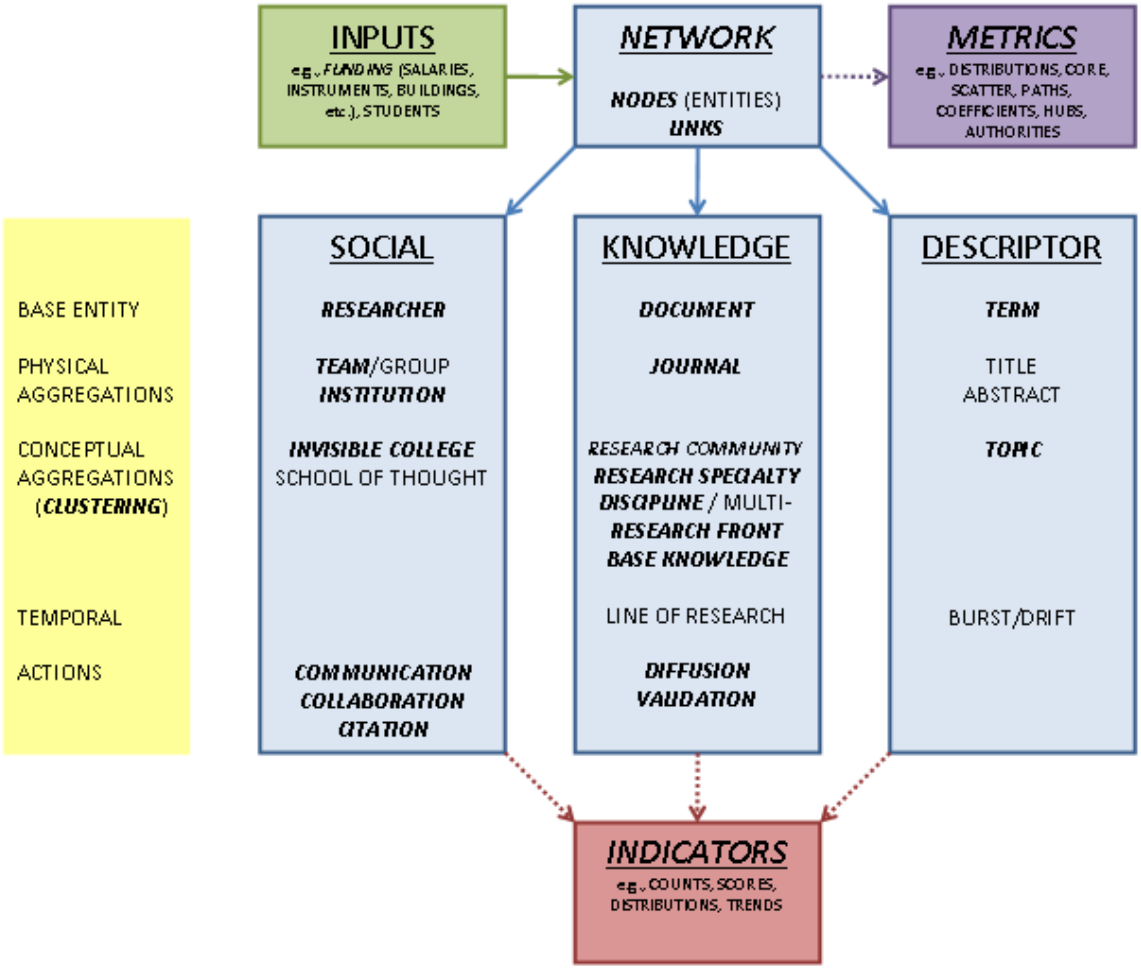

Figure 4. Network centric grouping of key terms used in science policy relevant modeling. Terms in bold italic are defined in the appendix.

4. Overview and Comparison of Major Science Models

This final section provides an overview and comparison of all major models captured in this book. 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 which development is closely connected to the process of diffusion of ideas" (Chapter 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 unsignificant" (Chapter 3, p.1);
- ***Stochastic models*** which are "appropriate when the system of interest is 'small' but when the fluctuations become significant for its evolution" (Chapter 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." (Chapter 4, p. 6)

- ***Evolutionary game theory (EGT)*** "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 (Chapter 5, p.2);
- ***Quantum game theory*** "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. Trough the concept of quantum entanglement, it is possible to include cooperative decision path caused by cultural or moral standards" (Chapter 5, p.18).

Theoretically, a reader might like to know what types of models exist and in what field of science they were developed (see Table 1, Columns 1-3). Interested in the coverage and applicability of the model, they might also like to know what data was used for validating the model (Column 6). Pragmatically, a reader might like to know exactly what questions a model aims to answer and what insights it provides (Columns 4-5).

Note that Chapters 3-5 each review one specific model type whereas Chapters 6-7 each discuss different types of models that address questions related to the structure and dynamics of co-author and paper-citation networks respectively.

Insert table ModelComparison-11.26.xls here

Table 1. Comparison of major science models covered in this book. For each model we list the model type (agent-based modeling (ABM), stochastic modeling (SM), the name of the model, the field of science in which the model was developed, the major questions the model aims to answer as well as key answers/insights in lay terms, the empirical dataset used for validation, citation reference, and chapter number in which the model is discussed.

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

Barabasi, A.-L., H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, 590-614.

Barabási, A. L., Reka Albert. (1999). Emergence of scaling in random networks. *Science, 286*, 509-512.

Ben-David, J., R. Collins. (1966). Social Factors in the Origins of a New Science: The Case of Psychology. *American Sociological Review, 34*(3), 451-465.

Bettencourt, Luís M.A., David I. Kaiser, Jasleen Kaur, Carlos Castillo-Chavez, David E. Wojick. (2008). Population Modeling of the Emergence and Development of Scientific Fields. *Scientometrics, 75*(3), 495-518.

Börner, Katy, Chaomei Chen, Kevin W. Boyack. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology (ARIST)* (Vol. 37, pp. 179-255). Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.

Börner, Katy, J.T. Maru, R.L. Goldstone. (2004). The simultaneous evolution of author and paper networks. *PNAS, 101, Suppl. 1*, 5266-5273.

Börner, Katy, Andrea Scharnhorst. (2009). Visual Conceptualizations and Models of Science. *Journal of Informetrics, 3*(3), 161-172.

Bradford, S. C. (1934). Sources of Information on Specific Subjects. *Engineering: An Illustrated Weekly, 137*, 85–86.

Callon, M., J. P. Courtial, W. Turner, S. Bauin. (1983). From Translations to Problematic Networks: An Introduction to Co-Word Analysis. *Social Science Information 22*, 191-235.

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago: University of Chicago Press.

Crane, D. . (1969). Social Structure in a Group of Scientists: ATest of the "Invisible College" Hypothesis. *American Sociological Review, 34*(3).

de Solla Price, Derek J. (1963). *Little Science, Big Science*. New York: Columbia University Press.

de Solla Price, Derek J. (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science, 27*, 292-306.

Elkana, Yehuda, Joshua Lederberg, Robert K. Merton, Arnold Thackray, Harriet Zuckerman. (1978). Introduction to "Toward a Metric of Science: The Advent of Science Indicators". In Yehuda Elkana, Joshua Lederberg, Robert K. Merton, Arnold Thackray & Harriet Zuckerman (Eds.), *Toward a Metric of Science: The Advent of Science Indicators*. Hoboken, NJ: John Wiley and Sons.

Farkas, I., I. Derenyi, H. Jeong, Z. Neda, Z. N. Oltvai, E. Ravasz, A. Schubert, A. -L. Barabási, T. Vicsek. (2002). Networks in life: Scaling properties and eigenvalue spectra. *Physica A, 314*(1-4), 25-34.

Gibbons, Michael, Camille Limoges, Helga Nowotny, Simon Schwartzman, Peter Scott, Martin Trow. (1994). *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*. New York: Sage Publications.

Goffman, W. (1966). Mathematical Approach to the Spread of Scientific Ideas: The History of Mast Cell Research. *Nature, 212*, 449-452.

Gupta, B. M., P. Sharma, C.R. Karisiddappa. (1997). Growth of Research Literature in Scientific Specialties. A Modeling Perspective. *Scientometrics, 40*, 507-528.

Kuhn, Thomas Samuel. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Latour, Bruno, Jonas Salk, Steve Woolgar. (1986). *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Leydesdorff, Loet. (1995). *The Challenge of Scientometrics: The development, measurement, and self-organization of scientific communications*. Leiden: DSWO/Leiden University Press.

Lotka, A. J. (1926). The Frequency Distribution of Scientific Productivity. *Journal of the WashingtonAcademy of Sciences, 16*, 317-323.

Lucio-Arias, Diana, Loet Leydesdorff. (2009). The Dynamics of Exchanges and References among Scientific Texts, and the Autopoiesis of Discursive Knowledge. *Journal of Informetrics, 3*(3), 261-271.

Merton, Robert. (1973). *The Sociology of Science*. Chicago: University of Chicago Press.

Morris, S. A., B.V. Martens. (2008). Mapping Research Specialties. *Annual Review of Information Science and Technology, 42*, 213-295.

Morris, S. A., G. Yen. (2004). Crossmaps: Visualization of Overlapping Relationships in Collections of Journal Papers. *Proceedings of the National Academy of Sciences of the United States, 101*(Suppl. 1), 5291-5296.

Mullins, N. C. (1973). *Theories and Theory Groups in Contemporary American Sociology*. New York: Harper and Row.

Mullins, N. C. . (1972). The Development of a Scientific Specialty: The Phage Group and the Origins of Molecular Biology. *Minerva, 10*, 52-82.

Nagurney, Anna. (1997). *Financial Networks: Statics and Dynamics*. Berlin; New York: Springer-Verlag.

Newman, M. E. J. (2001). Clustering and Preferential Atachment in Growing Networks. *Physical Review E, 64*(2), 025102-025104.

Newman, M. E. J., Leicht. E. A. (2007). Mixture Models and Exploratory Analysis in Networks. *Proceedings of the National Academy of Sciences of the USA, 104*, 9564-9569.

Nowotny, H., P. Scott, M. Gibbons. (2001). *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*. Cambridge, UK: Polity Press.

Rodriguez, M.A., J. Bollen, H. Van de Sompel. (2007). *A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and Their Usage.* Paper presented at the ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007, Vancouver, Canada, June 18-23, pp. 278-287.

Watts, D. J., S.H. Strogatz. (1998). Collective dynamics of "small-world" networks. *Nature, 393*, 440.

Whitley, R. . (2000). *The Intellectual and Social Organization of the Sciences*. Oxford: Oxford University Press.

Wojick, David E. , Walter L. Warnick, Bonnie C. Carroll, June Crowe. (2006). The Digital Road to Scientific Knowledge Diffusion: A Faster, Better Way to Scientific Progress? *D-Lib Magazine, 12*(6).

Zhang, H., B. Qiu, K. Ivanova. (2010). Locality and Attachedness-Based Temporal Social Network Growth Dynamics Analysis: A Case Study of Evolving Nanotechnology Scientific Collaboration Networks. *Journal of the American Society for Information Science and Technology, 61*, 964-977.

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison Wesley Press.

Zuccala, A. (2006). Modeling the invisible college. *Journal of the American Society for Information Science and Technology, 5*(2), 152-168.

Appendix

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 term that can be easily misunderstood. It has both noun and verb senses. The act of citation occurs when one **document** references another document. In a network sense, the citation can be thought of as a directed **link** between the citing and cited documents. Citations thus accrue to cited documents. In citation analysis, one speaks of a document having been cited n times, or having n citations. The word citation is often misused. For example, a reference is an item cited in a document – the bibliographic entry in a document. Yet, many refer to the reference as a citation. In addition, many others refer to a bibliographic record in a database as a citation. These latter two cases are improper uses of the word citation.

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 – variables, or according to the type of relationships to the other units – blockmodeling). 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 – From Wikipedia: "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." 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" (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 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, Warnick, Carroll, & Crowe, 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 show relationships between **terms**.

Network metric – A variety of metrics are used to characterize properties of networks. These include line (or keep 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 that performs research. In terms of modeling science, a researcher is not only one who performs researcher, 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 that apply for and receive funding through grant proposals.

Research community – Many years ago, sociologists, specifically Kuhn and Merton (1962; 1973) suggested that **researchers** organize into relatively small socio-cognitive groups – on the order of 10 people – working on common problems. Although the word community implies a measurement 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 five 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 (tens) research communities and is comprised of, on average, hundreds of articles per year. Lucio-Arias and Leydesdorff (2009) use "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** that 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 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.)