Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 75, No. 3 (2008) 495–518
DOI: 10.1007/s11192-007-1888-4

Population modeling of the emergence and development of scientific fields

LUÍS M. A. BETTENCOURT,[a,f] DAVID I. KAISER,[b] JASLEEN KAUR,[a,c]
CARLOS CASTILLO-CHÁVEZ,[d] DAVID E. WOJICK[e]

[a] *Theoretical Division, T-7 MS B284, Los Alamos National Laboratory, Los Alamos (USA)*
[b] *Center for Theoretical Physics, Laboratory for Nuclear Science, Department of Physics, Massachusetts Institute of Technology, Cambridge (USA)*
[c] *School of Informatics, Indiana University, Bloomington (USA)*
[d] *Department of Mathematics and Statistics, Arizona State University, Tempe (USA)*
[e] *Office of Scientific and Technical Information, US Department of Energy, Oak Ridge (USA)*
[f] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501 (USA)*

We analyze the temporal evolution of emerging fields within several scientific disciplines in terms of numbers of authors and publications. From bibliographic searches we construct databases of authors, papers, and their dates of publication. We show that the temporal development of each field, while different in detail, is well described by population contagion models, suitably adapted from epidemiology to reflect the dynamics of scientific interaction. Dynamical parameters are estimated and discussed to reflect fundamental characteristics of the field, such as time of apprenticeship and recruitment rate. We also show that fields are characterized by simple scaling laws relating numbers of new publications to new authors, with exponents that reflect increasing or decreasing returns in scientific productivity.

Introduction

Generations of scholars and science policymakers have chased an elusive goal: developing a science of science, some quantitative means of describing – and perhaps predicting – the growth and development of scientific research. For decades, scientists, historians, sociologists, bibliographers, and other researchers have sought some means of bringing order to the highly complex research enterprise. What factors might explain the changes over time of numbers of publications on a given topic, or numbers of

Received July 9, 2007

Address for correspondence:
LUÍS M. A. BETTENCOURT
Theoretical Division, T-7 MS B284, Los Alamos National Laboratory
Los Alamos, USA
E-mail: lmbett@lanl.gov

0138–9130/US $ 20.00

authors working in a given field? What interactions between researchers might account for a scientific community's growth and change? From the welter of potential influences on the scientific process, might some simplified, efficient model capture the bulk dynamics of how scientific fields emerge and develop?

Crucial steps were taken in the 1960s. Some pointed out the pervasiveness of logistic (or "S-shaped") curves [1]: whether measuring numbers of authors or numbers of publications, many fields began with a burst of exponential growth followed by saturation and an eventual steady-state. (The same pattern held for many other features of science, ranging from the number of known chemical elements over time to the energies achieved by particle accelerators.) The ubiquity of such logistic curves, and their repeated appearance from the age of Galileo and Newton to the present day, seemed to point to some basic underlying structures of science (see also [2]).

Around the same time, other scholars began importing tools from epidemiology to study the spread of scientific ideas [3, 4]. (For a review, see [5].) Much as a virus spreads via contact from person-to-person, throughout a susceptible population, so too do novel ideas "infect" researchers. The simplest formulation, involving three classes of people (susceptibles, infectives, and those who have recovered), moreover, could reproduce logistic curves for such time series as numbers of publications or authors [3, 6]. More recently, others have focused on the structure and evolution of networks of co-authorship and citation (see, e.g., [7–11]). Together these types of studies provide much-needed statistical structural analyses of how scientific fields emerge and change over time.

Since these pioneering studies were published, new tools and resources have become available. Digital libraries and archives, in tandem with efficient search engines and the computational power to retrieve and parse massive amounts of information, make it practicable to expand the repertoire of models, and to test them against empirical datasets far larger than those considered in the original studies.

Building on many of these insights, we have developed a coarse-grained approach to modeling the time-evolution of scientific fields mathematically. Like earlier efforts, our model is inspired by epidemic contact processes suitably adapted to take into account the nature of social interactions and dynamical processes by which scientific ideas spread – social interactions gleaned from close empirical study of historical cases [12]. Variations in the small number of parameters can increase or hamper the speed at which a field develops. Moreover, we have tested our simple model against data from six separate emergent scientific fields, covering a broad range of disciplines, from physical sciences to medical research to cutting-edge technology. Some fields are essentially theoretical in nature, others more exclusively experimental. Many show signs of the tell-tale logistic curve, while others do not. Yet in each case, our parsimonious model produces extremely good fits to the data.

We have also pursued a simple measure of a field's productivity. Strict scaling laws appear to hold for each of the six scientific fields studied here, relating the number of new publications on a given topic over some time period (say a year) to the corresponding number of new authors entering the field during that time. Substituting number of new authors for the ordinary time variable thus reveals a universal underlying similarity in structure across these disparate fields, even though they betray different behavior when plotting either authors or publications versus time. Even those fields that depart from the familiar logistic curve (when plotted with respect to ordinary time) nonetheless obey the same kind of simple scaling law as those for which the logistic curve does hold. These scaling relationships, which are analogous to measures of productivity in economics, might therefore be a useful measure of a field's scientific health, measured by how publication rates vary with the addition of new authors.

The remainder of this paper is organized as follows. The methods section introduces and discusses our methods, which comprise of mathematical models to describe the population dynamics (numbers of authors) behind the establishment of scientific fields. The approach is based on a succinct (coarse) description of contact processes between scientists. We selected this model – a simplified version of a general class of models we have developed [12] – based primarily on its ability to treat a wide range of data patterns efficiently, across several different scientific fields. We also describe our methods for estimating parameter values, our optimization techniques used to match the model to data, and our method of generating error estimates. Finally we describe in detail the parameterization of scaling laws relating change in numbers of new papers to number of authors entering the field, and place these measures in the context of analogous measures of productivity used in economics. The results section presents our results and includes brief accounts of six case studies of scientific evolution, measured by the growth in number of active authors over time, together with the results of fitting our model to these data, including extrapolations to the near future. We also discuss the productivity structure of each of these fields. The discussion and conclusions section discusses these results and provides some perspectives on the values and limitations of the model. We also discuss topics for further research.

Methods

Data searches and time series construction

Time series data for six fields detailed below were assembled from keyword and citation searches using SearchPlus, which was developed by the Los Alamos National Laboratory's Research Library and Library Without Walls [13]. It searches an integrated set of the large scientific publishing databases, including BIOSIS®,

Engineering Index, Inspec, and ISI databases (Thomson Scientific), such as ISI Proceedings, ISI SciSearch, ISI Social SciSearch, ISI Arts & Humanities.

Results for each field, specified by publication title, author names, and publication reference (including publication year), were stored in relational databases after parsing to eliminate repeats and perform author-name matching. Cumulative and yearly differential numbers of unique authors and publications were then extracted and organized as time series for modeling and statistical analysis. Detailed keyword and citation searches for each field are given in Appendix A.

Population models of scientific dynamics

Our starting point is a generalized SEIR epidemic model [14]. In addition to the familiar susceptible (S), infected (I), and recovered (R) classes, we incorporate an "exposed" class (E): people who have been exposed to the new idea, but who do not yet manifest it in their published research. We added this new feature in the course of our previous study of the dynamics of how Feynman diagrams spread among communities of theoretical physicists [12]. In effect, it takes into account the crucial stage of training and apprenticeship: no scientist makes the leap from student to practitioner instantaneously, nor do professional scientists leap effortlessly from one specialty to another. The model also includes (exponential) population growth (recruitment) and multiple contacts between members of the exposed and infected classes, terms reflecting the social dynamics of scientific activity that likewise proved important in our previous study [12]. The model is written explicitly as

$$\begin{aligned} \frac{dS}{dt} &= \Lambda N - \beta S\frac{I}{N}, \quad \frac{dE}{dt} = \beta S\frac{I}{N} - \kappa E - \rho E\frac{I}{N}, \\ \frac{dI}{dt} &= \kappa E + \rho E\frac{I}{N} - \gamma I, \quad \frac{dR}{dt} = \gamma I, \end{aligned} \tag{1}$$

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 (that is, those who have adopted the new scientific idea, as manifested in their publications), and $R(t)$ is the size of the population who have recovered (no longer publishing on the topic). We shall refer to the size of the entire population, the sum over these classes, as N: $N = S+E+I+R$. Note that we did not include an exit (or death) term, as this tends to be very small, and is 'subsumed' by the recovered class. In this model, the population, N, grows exponentially with rate Λ. In some instances, indicated explicitly below, the growth term will be written as Λ, instead of ΛN, and may not apply to the entire duration of the dynamics.

The remaining parameters account for the probability and effectiveness of a contact with an adopter, β; the standard latency time, $1/\kappa$ (which, in this case, gives the average

duration of apprenticeship, after one has been exposed but before one manifests the new idea in publication); the duration of the infectious period, $1/\gamma$ (how long one publishes on the topic and can teach others); and the probability that an exposed person has multiple effective contacts with other adopters, ρ. The model may be visualized as in Figure 1. The reproductive number for the model – that is, the average number of new people infected by a given infected individual – is $R_0 = \beta/\gamma$. This is a simplified version of a more general family of models developed in the course of this work, which feature multiple latency classes [14].

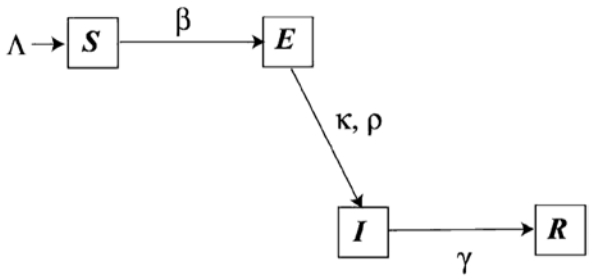


Figure 1. Flow diagram for the SEIR model

As illustrated below, 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 will grow essentially as a logistic curve. Increase the contact rate (β) or boost recruitment (Λ), however, and $I(t)$ will grow more nearly linearly, as indeed has been found empirically for some fields.[1]

Parameter estimation

Parameter estimation for our model is performed via a stochastic ensemble algorithm as described in [12]. The parameters describing the initial conditions [$S(t_0)$, $E(t_0)$, $I(t_0)$, $R(t_0)$] and the dynamical parameters (Λ, β, κ, γ and ρ) are organized as a vector of real-valued numbers. An ensemble of such vectors, or strings, is generated via the perturbation of a "progenitor" string. The fitness of the resulting strings – referred to collectively as a "generation," given the similarity to genetic algorithms – is evaluated by comparison with the data. A set of 'best' strings, in terms of fitting the data, is then

[1] Other epidemiological models, such as those developed to model the spread of sexually-transmitted diseases [15], also predict linear rather than exponential growth of the infected population when the diseases have long incubation times and people have multiple partners and long-term associations – not unlike the situation in the spread of scientific ideas.

chosen to spawn the subsequent generation, and so on, until the procedure converges, that is, the fitness ceases to improve. Several checks are performed to guarantee that the absolute (global) best fit was reached. The procedure generates not only a best fit solution, showing the smallest deviation to the data, but also an ensemble of good strings, which fit the data up to some user specified tolerance (here 10% variation per point in numbers of authors), from which uncertainty in the solution is quantified.

Below we apply this stochastic ensemble optimization procedure to data on the number of authors participating in the advent and subsequent growth of several fields, in the aftermath of a discovery, breakthrough, or surge of interest.

Scaling laws for scientific productivity

Population models focus usually on the number of *people* manifesting a certain feature in a given population. But scientific fields can also be categorized usefully based on their research output – that is, the number of publications on a given topic, rather than the number of dedicated researchers. Thus we have also examined the relationship between number of publications and number of authors as a given field evolves. We observed that all six cases obey a remarkable scaling law: yearly numbers of new publications scale as a simple power law with the corresponding number of new authors, which we write as:

$$\Delta \text{ Publications} = C(\Delta \text{ Authors})^{\alpha} . \qquad (2)$$

Here Δ denotes new publications or authors over some time period (that we will adopt as one year), C is a normalization constant, and α is the scaling exponent.

As we demonstrate below, Eq. (2) provides an excellent fit to data for all six fields, but with different values of the scaling exponent α. Note that for $\alpha > 1$ a field would grow by manifesting increasing returns to scale; specifically by showing an increase in the number of publications per capita. Thus $\alpha > 1$ characterizes a field with increasing individual productivity as a field attracts new scientists, which is a sign of opening new opportunities and vitality. Conversely a field characterized by $\alpha < 1$ shows per capita decrease in productivity as it develops and typically signals closing opportunity and a dying subject matter, where new papers require ever greater effort in terms of numbers of workers in the field. We show below that both cases characterize specific new fields of research and that there can be transitions between different productivity regimes, signaling such events as major shifts in funding or the occurrence of scientific breakthroughs.

Results

Case studies and population modeling results

In this section we present several studies of the emergence of scientific fields and results of modeling the evolution in numbers of authors using the model introduced in the Methods section. Parameter estimates are given at the end of the section for all cases.

Cosmological inflation

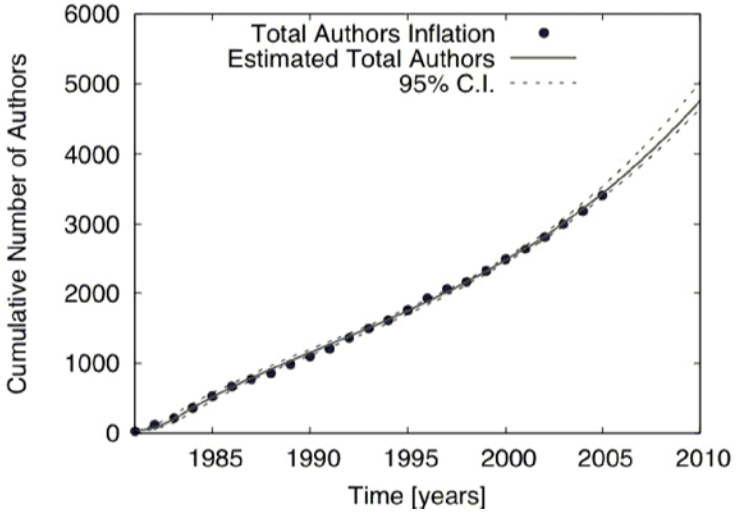
Cosmological inflation, proposed by Alan Guth in 1981 and quickly elaborated upon by others [16], describes the exponential expansion of the volume of spacetime during the early universe. Remarkably, it provides solutions to most open issues that arise when combining big bang cosmology with astrophysical observations. Originally it was conceived to solve the problem of overproduction of cosmological defects (mostly monopoles, see cosmic strings below), but perhaps its strongest feature is to provide a prediction for the initial energy density perturbations necessary to seed the large-scale structure of the universe [17].

The data shown here result from literature searches based on citations to an early set of publications in the field as well as to later review articles (see Appendix A). The time evolution of the field, measured in terms of numbers of authors, is peculiar when compared to other case studies discussed below: it is approximately linear, appearing least like a classic logistic S-curve. Nevertheless, our SEIR model with population growth and large contact rate describes the data very well (see Figure 2a).

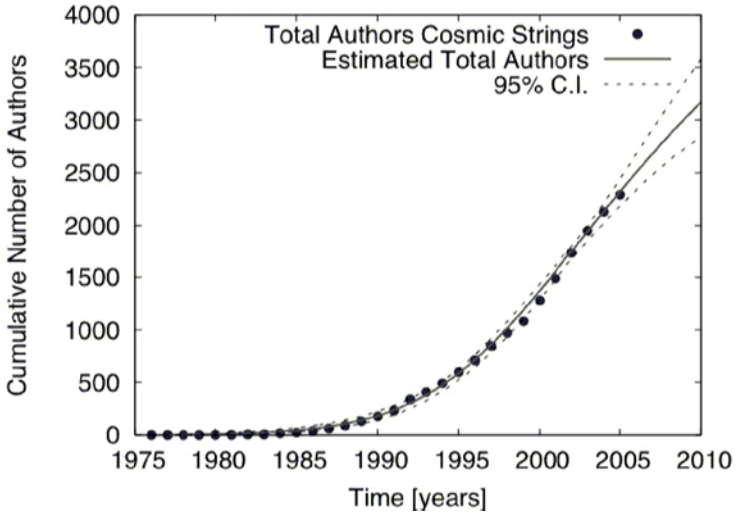
Cosmic strings

Cosmic strings and other topological defects are non-perturbative solutions of unified theories of elementary particle interactions that may have been formed in phase transitions in the early universe. In 1976 T. W. B. Kibble [18] suggested that the internal symmetry groups of modern theories of particle physics – taken together with the Higgs mechanism, necessary to render elementary particles massive – lead inexorably to phase transitions in the early universe, similar to, but generally more complex than those in superconductors and certain superfluids. It follows that, in the early universe, as in these materials, topological defects could form that would be stable and able to concentrate vast amounts of energy in their profiles, be they monopoles, cosmic strings, or domain walls. These defects could seed the large-scale structure of the universe by providing energy (and momentum) inhomogeneities upon which baryonic matter could fall [19].

For some time, until the late 1990s, Cosmic Strings and Cosmological Inflation were rival theories contending to explain the features of the observed universe.



a)



b)

Figure 2. The temporal evolution of the cumulative number of authors (dots) publishing in a) Cosmological Inflation and b) Cosmic Strings, the fit from the model (solid line), and 95% confidence interval (dashed lines)

Both fields are similar sociologically, involving mostly theoretical work with high energy physics models, and the two author communities often mixed: approximately 15% of all the authors who published on either Inflation or Cosmic Strings also published on the other field, as revealed by the author pools in each dataset. For these

reasons it is very interesting to compare the temporal development of the two fields. Each field is equally well fit by the model of Eq. 1, although with rather different values for the parameters (see Figure 2 and Table 1). In particular, whereas Inflation deviates sharply from the familiar logistic curve, Cosmic Strings follows this trajectory closely. Although both fields grew rapidly during the 1980s and 1990s, recent precise measurements of the cosmic microwave background and type Ia supernovae seem to weigh more strongly in favor of Inflation, making the case for cosmological defects increasingly constrained [17]. This most likely accounts for the up-turn in numbers of authors publishing on Inflation and the declining rate of new authors pursuing Cosmic Strings. In spite of these trends, however, the field remains fast expanding with hundreds of publications and new authors every year.

Prions

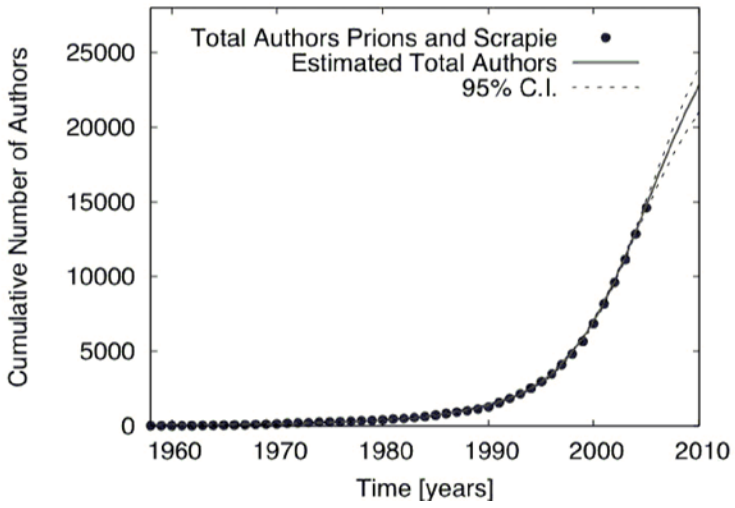
In addition to the theoretical research of the previous two examples, we also sought to characterize two examples of research in biology and medicine, namely Prions and Scrapie and the more recent explosion of interest in avian influenza or "bird flu," more technically known as H5N1 influenza.

Prions (proteinaceous infectious particle) are abnormally configured proteins, which were shown in 1982, by S. B. Prusiner and colleagues [20], to cause scrapie, a transmissible spongiform encephalopathy in sheep. It was later recognized that other related diseases, such as Kuru (Creuzfeld-Jacob disease in humans) and BSE ("mad cow disease"), are also caused by prions, and not by a virus or any other conventional infectious agent. The discovery was followed, a decade later, by great public health scares (and interest), principally associated with BSE in the UK, which contributed to raise the profile of the field. By the late 1990s research in prions had become underfunded, and the field started to show signs of slow down [21]. For the discovery of prions and their connection to spongiform encephalopathies, Stanley B. Prusiner won the Nobel prize in Medicine in 1997 [22].

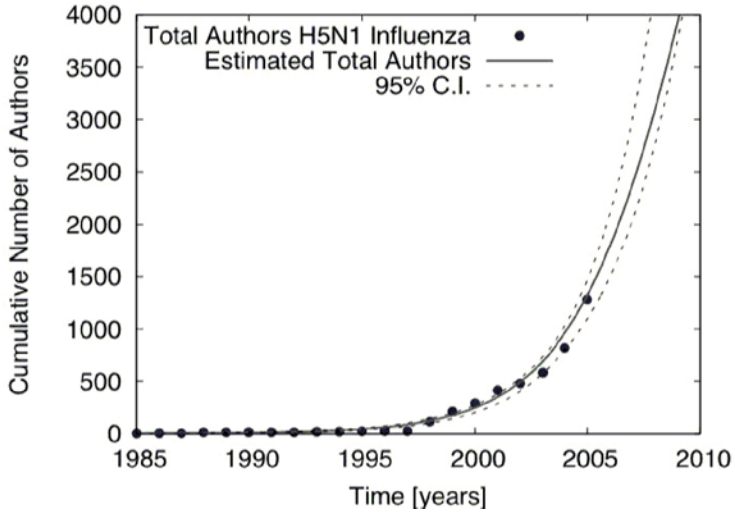
The data used here were obtained from a keyword search for "prion" among scientific publications (having eliminated a genus of like-named birds), and also includes research in scrapie (see Appendix A). The growth curve of numbers of authors in Prions and Scapie – similar in shape to that for Cosmic Strings but larger by about one order of magnitude – closely resembles the familiar S-shape (see Figure 3a). The field potentially shows signs of saturation, as discussed below (see Table 1).

H5N1 influenza

Unlike prions, research in H5N1 influenza is a very young field, driven by the health emergency of a possible new devastating influenza pandemic. H5N1 influenza is a subtype of the influenza A virus that causes "bird flu." It is presently a disease of birds, but, as of July 2007, there have been over 280 confirmed cases of human infection, mostly in Southeast Asia, with an observed case mortality rate of about 57% [23].



a)



b)

Figure 3. The temporal evolution of the cumulative number of authors publishing in a) Prions and Scrapie and b) H5N1 influenza (dots), the fit from the model (solid line), and 95% confidence interval (dashed lines)

Research in H5N1 and other types of influenza has gained extraordinary impetus over the last few years. Research on the H5N1 subtype started in earnest after 1997, when the first human cases of the disease were identified in Hong Kong. New, larger outbreaks in the last 2-3 years have lead to great scientific and public interest in the

field and its relation to other influenza types. Its literature spans themes in health policy, epidemiology, and molecular biology (see Appendix A).

The evolution of the number of authors, while still small compared to our other examples, is fast increasing and shows no signs of saturation, as reflected in our model fit (see Figure 3b and Table 1).

Carbon nanotubes

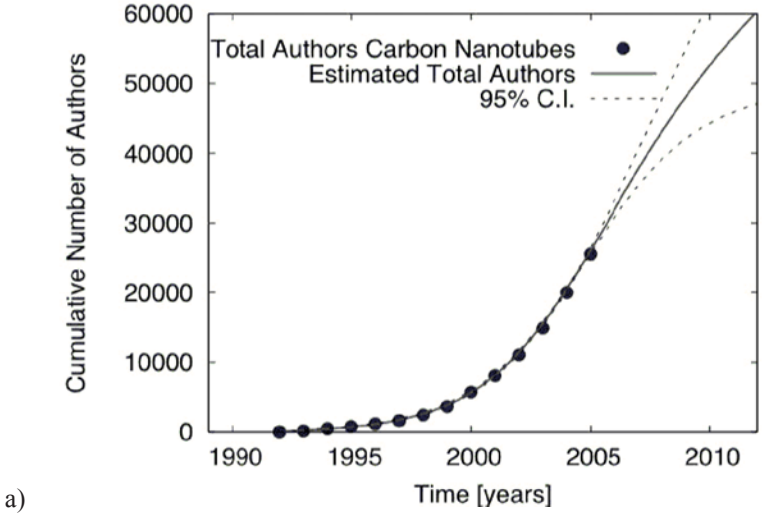
We also sought to include case studies in fields with a strong technological component, which nevertheless also contain important theoretical contributions. Carbon nanotubes is one of the more tangible subfields of nantechnology and combines strong elements of new technologies and materials science theory. Quantum computer and computation (see below) similarly spans recent discoveries in computer science, quantum engineering, and nanoscale devices.

Carbon nanotubes are a recently discovered allotrope of carbon, which promises to generate a whole family of new materials and potentially revolutionize nano-engineering. Research in this area started in 1991, when Sumio Iijima of NEC in Japan discovered a new method (arc discharge) to produce them, although nanotubes had been described before in the literature [24]. It is hoped that these materials may usher in many promising engineering solutions at the nanoscale due to their enormous strength, lightness, and conductive properties of heat and electrical currents. Hence the scientific literature in this and closely related areas has grown rapidly during the past two decades. [6, 25, 26]

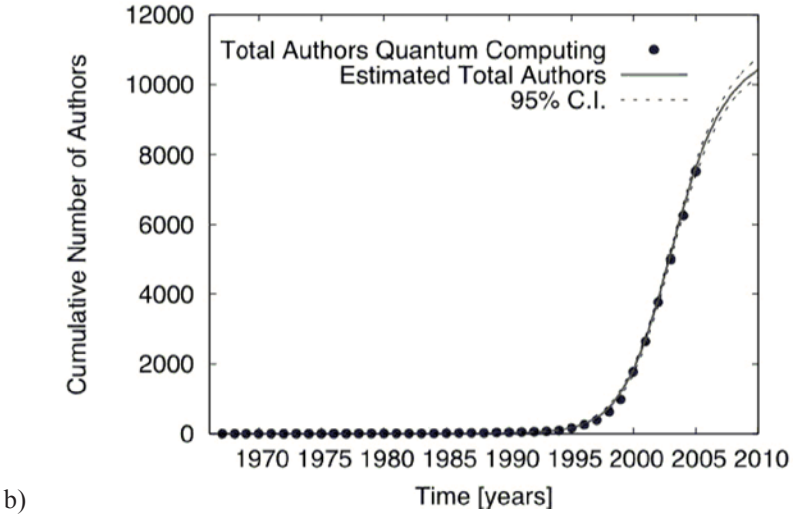
We built a database of scientific publications in the field via keyword searches (see Appendix A). Although the field is still young it has the greatest number of authors among our examples. It shows robust and fast growth, consistent with earlier findings [25] that showed nanotubes to be the fastest-growing field within nanoscience and technology. The growth of research on carbon nanotubes is well fit by our model. Time will tell whether or not research in this area will begin to saturate over the next decade or so, as predicted by our model; other areas within nanotechnology, such as fullerenes, have indeed already shown signs of such saturation [6] (see Figure 4a and Table 1).

Quantum computing and computation

Quantum computing is an emerging field of research dedicated to the discovery of new devices and theoretical implementations of states that are genuinely quantum-mechanical and can be manipulated for computation. A quantum computer would be able to perform certain operations (such as factorization) much faster and solve physical quantum models more naturally and efficiently than any classical computer. The field is naturally multidisciplinary, involving research in quantum theory, computer science, materials science, and engineering.



a)



b)

Figure 4. The temporal evolution of the cumulative number of authors (dots) publishing in a) Carbon Nanotubes and b) Quantum Computing and Computation, the fit from the model (solid line), and 95% confidence interval (dashed lines)

The first references to quantum computation date back to the late 1960s and 1970s, but the subject only gained momentum with several theoretical breakthroughs in the 1980s and 1990s. In 1982 Richard Feynman [27] showed how quantum systems could be used to do computation. In 1985 David Deutch [28] proved that any physical process

can, in principle, be modeled perfectly in a quantum computer. In the mid 1990s algorithms by Shor and by Grover [29] and NMR experiments [30] demonstrated explicitly the advantage of quantum algorithms in certain computational tasks and the first implementations of the technology. These discoveries opened up the field to wide expertise and changed its recruiting dynamics, as we discuss below in more detail.

Figure 4b shows the temporal evolution in numbers of authors in Quantum Computers and Computation. After a long simmering period of general interest the field took off in earnest in the mid-1990s and has shown fast growth since. These trends are very well fit by our model, which also predicts the onset of some saturation in the time frame of 5-10 years.

Parameter estimates

Table 1 shows the summary of parameter estimates for all six emergent scientific fields. Our optimization method allows us to estimate both initial populations of susceptibles, exposed, adopters (infectious), and recovered as well as the dynamical model parameters. In all cases, the recovered population at the initial time, $R(t_0)$, was negligible. The number of initial exposed, $E(t_0)$, is also small in most cases, with the exception of Carbon Nanotubes, where many scientists in the field were working with other similar allotropes of carbon (such as other fullerenes) and were essentially ready to contribute to the field. The initial population of susceptibles, $S(t_0)$, in many cases reflects the field's size, as does its recruitment rate, Λ. Some fields, such as Cosmic Strings, Prions, and H5N1 influenza are better fit by a constant recruitment rate, independent of the scientific community size, of the order of a few hundred researchers a year. Other fields, such as Cosmological Inflation, Carbon Nanotubes, and Quantum Computing are better characterized by early recruitment rates that are proportional to population, varying between less than 10% for inflation to 40–50% in the two technological fields, indicating a much larger rate of infusion of new researchers. We consistently found that contacts between exposed and infectious populations (denoted by ρ) are not necessary to provide a good description of the dynamics. The duration of the typical incubation time, $1/\kappa$, before an exposed individual becomes infectious was roughly consistent across fields and varied between 1.4–5 years, which is a reasonably average apprenticeship time for new researchers (e.g. graduate students or postdoctoral fellows). The duration of the time over which an individual can transmit the idea, $1/\gamma$, varied more widely, between about 6 months (Cosmic Strings) to 10 years (Carbon Nanotubes), suggesting a much larger turnover time for researchers in some fields than in others. Finally the reproductive number for each field, R_0, which measures the average number of susceptibles that an idea adopter infects, is always large compared to similar ratios for infectious diseases and varied between about 2–65.

These large numbers are typically not the result of large contact rates β (with the possible exception of Cosmological Inflation) but rather of long infectious periods that allow an idea to be slowly developed over the period of several years and transmitted many times.

Table 1. Parameter estimates for the model of Methods section, and data sets described in Results section

Parameter	Cosmological inflation	Cosmic strings	Prions & scrapie	H5N1 influenza	Carbon nanotubes	Quantum computing
$S(t_0)$ *	930±1	14±9	14262±1368	9057±200	30464±5976	11627±91
$E(t_0)$	6	5	1	1	501±24	0
$I(t_0)$	37±1	0	8±1	0	1	0
$R(t_0)$	2	0	7±2	0	1	0
β	13.41±0.28	4.45±0.42	0.69±0.05	1.47±0.02	0.99±0.05	3.78±0.09
Λ	0.07	159.1±2.7*	469±25*	138±10*	0.04±0.01	1.03±0.02**
κ	0.20	0.25±0.02	0.22±0.01	0.71±0.01	0.50±0.03	0.41±0.02
ρ	0	0	18.4±1.24	0	0.03±0.06	0.77±0.03
γ	0.21	1.73±0.19	0.37±0.03	0.6±0.01	0.10±0.05	1.18±0.02
R_0	64. ±1.5	2.58±0.11	1.87±0.03	2.44±0.03	9.72±1.71	3.20±0.11
α	1.28	1.13	0.78	0.87	1.32	1.37***

* Indicates a linear growth term Λ, rather than ΛN in the equations for S.
** Indicates that the susceptible population growth starts in 1990.
*** This value for α applies only once the number of new authors reaches ca. 1000 in a given year; for smaller new-author pools, the best fit productivity curve yields $\alpha = 1.00$.

Scaling and scientific productivity

Having modeled the evolution of numbers of authors we may now ask how these dynamics relate to the overall productivity of the field, at least as measured by the number of publications. Note that we present results for numbers of new publications and authors over some time period, but have not made any attempt to distinguish between impact factors characterizing different researchers or publications.

We found extraordinary consistency when analyzing how growth in numbers of publications relates to numbers of new authors. For all six cases, the scaling law of Eq. (2) fits the data very well, regardless of the details of the dynamics described above (see Figures 5–7). Exponents α, which characterize each field's productivity, did vary, however, showing increasing returns to scale ($\alpha > 1$) for the theoretical and technological fields (with the latter showing largest α), and decreasing returns ($\alpha = 0.8$) for the biological and medical fields. Moreover the field of Quantum Computation shows a clear transition, around 1994–1995, between a period in which motivation existed for research in the area but no tangible technical breakthroughs had been made

($\alpha = 1$), to a period of large increasing returns to scale, after new experimental and algorithmic paths had been identified. This sharp shift in α is consistent with our population modeling, which produced the best fit when the field's recruitment, Λ, was "switched on" in 1990.

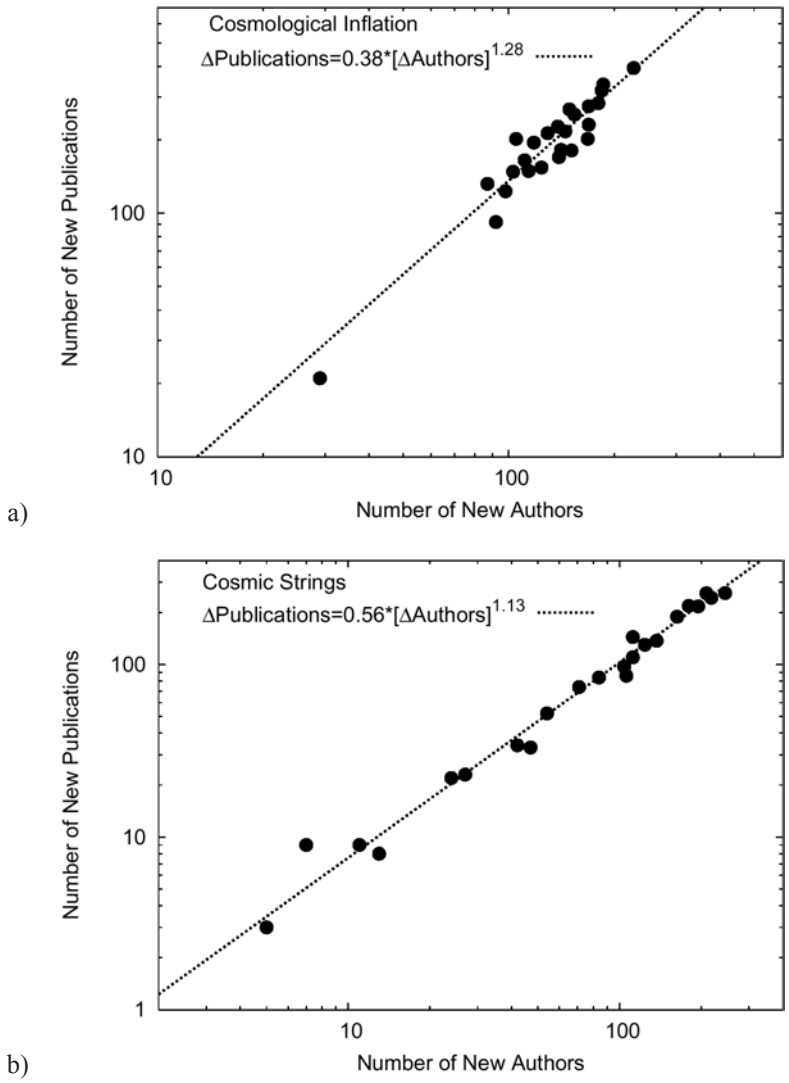
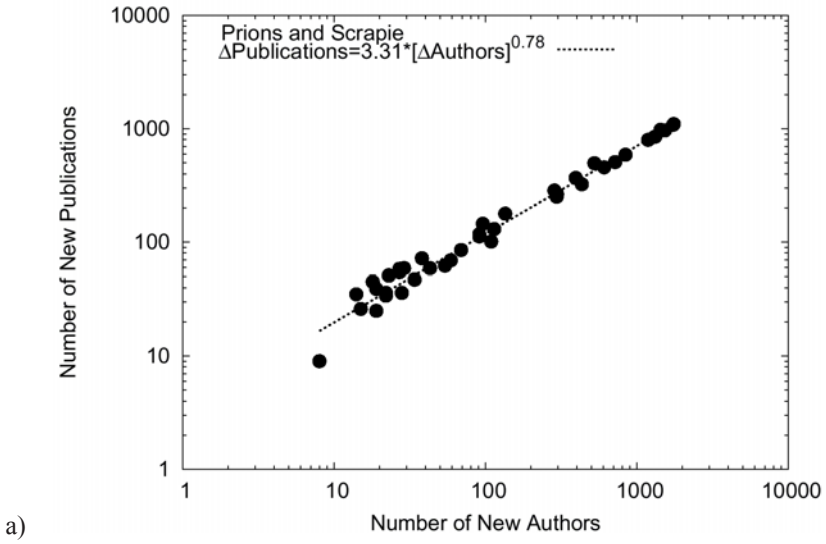
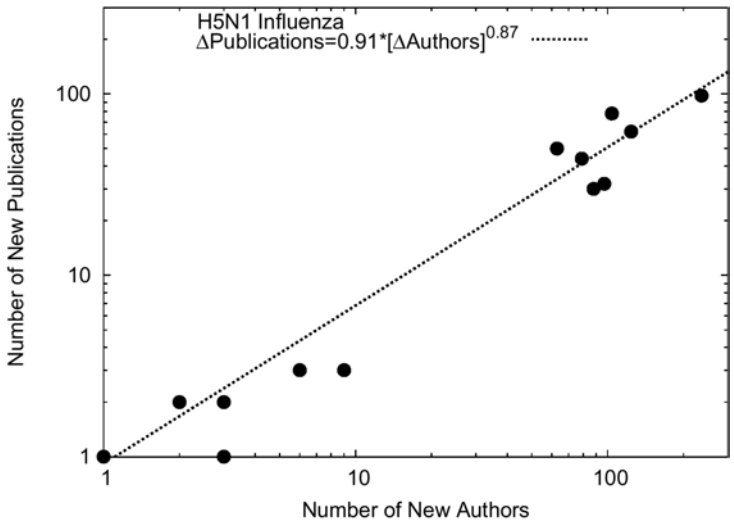


Figure 5. Productivity curve for research on a) Cosmological Inflation and b) Cosmic Strings



a)



b)

Figure 6. Productivity curve for research on a) Prions and Scrapie and b) H5N1 Influenza

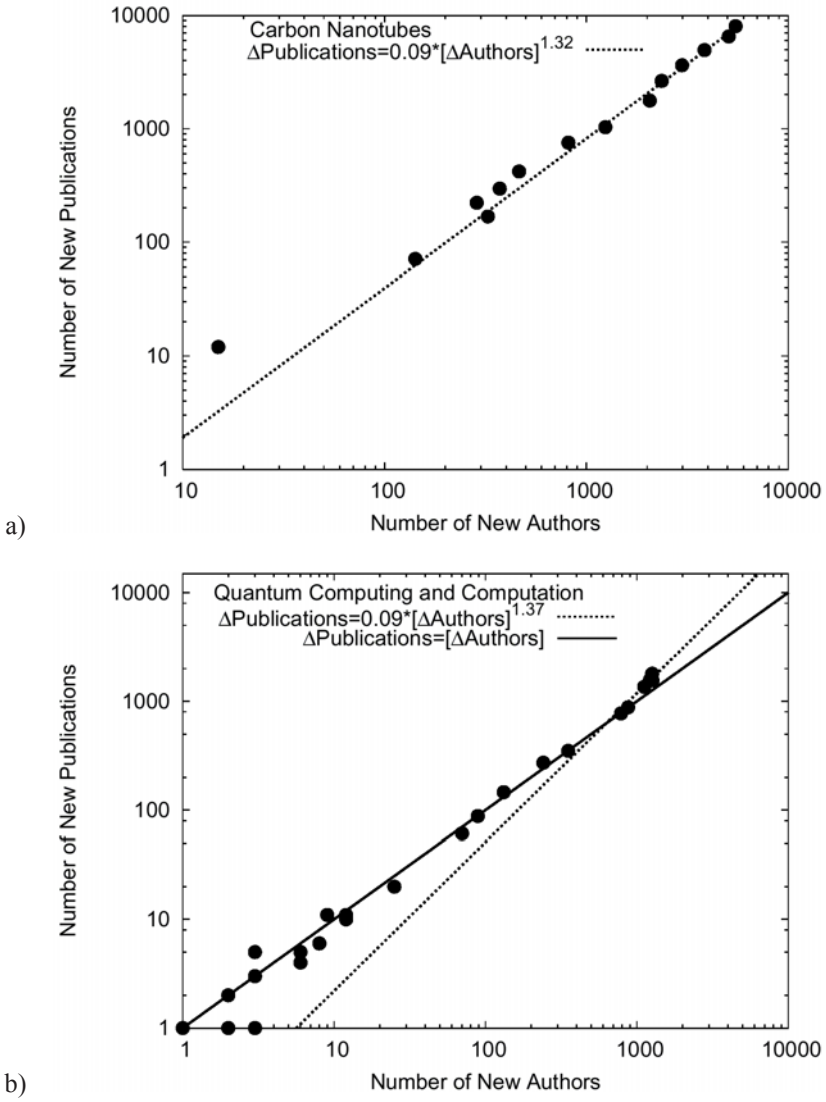


Figure 7. Productivity curve for research on a) Carbon Nanotubes and b) Quantum Computing and Computation

Discussion and conclusions

The six case studies analyzed here show that population models analogous to those of epidemiology, suitably adapted, provide excellent bases with which to describe

quantitatively the emergence and development of scientific fields across the natural sciences. Remarkably, our simple model describes equally well theoretical fields (such as Cosmological Inflation and Cosmic Strings) as experimental ones (such as Prions and Carbon Nanotubes), or those that include both kinds of activity (such as Quantum Computing and H5N1 influenza). This constitutes an important demonstration that the type of model we devised to treat one particular case in great detail (the spread of Feynman diagrams among theoretical physicists, [12]) may be applied much more broadly, with equally impressive fits to empirical data.

Moreover, the parameter estimates for these many cases reveal several features that make intuitive sense. For example, the two cases that draw authors from only one narrowly defined specialty – Cosmological Inflation and Cosmic Strings – show the smallest initial populations of susceptibles, $S(t_0)$, whereas those fields that cross disciplinary boundaries, potentially attracting researchers from many different scientific areas, reveal correspondingly larger initial populations of susceptibles. Likewise, the two purely theoretical fields (again, Cosmological Inflation and Cosmic Strings) show the greatest effectiveness of contact, β; those fields which most thoroughly mix theoretical and experimental components (Quantum Computing and H5N1 influenza) have intermediate values for β; while the more purely experimental fields (Prions and Carbon Nanotubes) have the smallest values for β. As one might expect, it is one thing to practice and master a pencil-and-paper technique; quite another to build and oversee an entire working lab group.

We must note however that although the fundamental dynamics of contact and spread may be analogous between the spread of ideas and disease, many characteristics of the two dynamics are fundamentally different. First, the nature of the contacts is clearly distinct. Many scientific contacts are prolonged and based on mentor/apprenticeship relationships such as those between advisors and students or postdocs. This fact also highlights that recruitment plays an important role, alongside conversion of susceptibles in the growth of a field. Parameter estimation supports these expectations, showing large numbers of initial susceptibles and/or population growth typically of a few percent a year.

Compared to most diseases, scientific ideas spread slowly, taking years to become adopted by a significant number of practitioners. They also show substantial contact rates over these time scales, perhaps the result of the many intentional social structures – PhD programs, postdocs, meetings, workshops, etc. – designed to foster sustained interactions. The result is that although incubation times range between 1.4–5 years and infectious periods between 6 months and 10 years, all cases show large basic reproductive numbers R_0 between 1.8 and 64. This point seems to be very general and a manifestly different feature between ideas and infectious diseases: useful ideas may never be forgotten, leading to very long infectious periods and therefore large R_0.

It is also typical of the spread of ideas that long and repeated contacts between adopters and susceptible individuals take place in order for the concept or technique to be transmitted. Here we modeled these processes via a contact term between exposed and infected, proportional to the contact rate ρ. In most of our case studies, however, with the exception of Prions & Scrapie, estimation of this term shows very small values for ρ, indicating that perhaps persistent contacts were not essential, or that a different modeling strategy might be necessary to capture such effects.

While most case studies showed growth dynamics that are familiar from other invasion processes, two of our examples were peculiarly different. First, Cosmological Inflation shows growth in numbers of authors that has been remarkably linear over more than 20 years, without displaying the more typical phase of exponential growth. Nevertheless, our model provides an excellent fit to the data, even though parameter estimates force the numbers of exposed and infected to their fixed points, as functions of a growing population N, at a relative growth rate of 7% a year. As a result the model solution that best fits the data for Inflation is particularly sensitive to the growth dynamics of the population of susceptibles, and less so to the magnitude of the contact rate, as this factors out in the fixed point solution for $I(t)$. It would be interesting to validate these inferences, or seek good fit solutions in different regimes.

Quantum Computing and Computation shows a particularly long incipient period, with very slow growth over more than twenty years, and a quick (approximately exponential) rise starting in the late 1980s. We modeled these dynamics by allowing for susceptible population growth starting only in 1990, which gives an excellent fit to the data. We note, however, that models with several exposed classes and therefore potentially longer successive incubation times, or with time-varying contact rates, may also provide viable alternatives. As for the case of Inflation, it would be interesting to obtain more detailed historical data that could guide such detailed modeling choices.

The type of modeling described here can be enlarged in several interesting directions. One direction involves improvements to the model itself. The distribution of the length of the infectious period, recruitment rates, and perhaps even incubation times may be inferred directly from publication data, PhD theses records, and so on. Knowledge of their distributions would help greatly to constrain and improve models, as well as distinguish whether the growth of a field is primarily the result of the recruitment of new susceptibles, or instead the consequence of the conversion of an already large susceptible population via a larger contact rate. Additional features of the basic SEIR model may also be added, such as a model in which the size of the infected class facilitates further recruitment (directly linking I with Λ). Similarly, an explicit class of converts to competing ideas, Z ("skeptics" or "stiflers"), may be added, as in [12]. This feature could prove especially useful for the combined modeling of Cosmological Inflation and Cosmic Strings, which were competing fields where nevertheless many authors published on both topics over time.

Finally we observe that while the dynamics in terms of numbers of authors differs from field to field, the relation between numbers of new publications to that of new authors appears remarkably simple, following in each case the simple scaling law of Eq. (2). This suggests a self-similarity of dynamics that is characteristic of each field, suggesting that recruitment (of susceptibles via author pool growth) is the fundamental driver of scientific development, with productivity per author of specific fields being stable even as the field grows in size by many orders of magnitude. The exponents α denote this measure of productivity, showing in our analyses increasing returns to scale, $\alpha > 1$ (i.e. increasing number of papers per capita) as a field grows in most studied cases, and decreasing returns, $\alpha < 1$, in others. This type of measure of productivity is ubiquitous in other socio-economic systems, where its dynamics may itself drive growth. In that case it has been shown [31] that dynamics under decreasing returns always asymptotes to a finite size population of authors and publications, while those under increasing returns may lead to indefinite growth, and show clear growth cycles. The existence of these dynamics in scientific literatures is an interesting question for further research.

Specifically, in terms of productivity, the six cases fall out into fairly neat clusters: nanotechnology fields (Carbon Nanotubes and Quantum Computing) show the greatest increase of publications versus authors, followed by theoretical physics topics (Cosmological Inflation and Cosmic Strings), followed by more applied biomedical research (Prions, H5N1 influenza). One might have naively expected the order to be slightly different – with the theoretical physics fields showing greatest exponents, followed by nanotech – on the idea that it costs very little to set up a new research group and get them up to speed on theoretical topics, for which no equipment and little infrastructure is needed. Especially given the recent change in slope for Quantum Computing (from $\alpha = 1$ to $\alpha = 1.37$), these large values of α in the nanotechnology fields might well reflect the intense funding and media attention granted to such areas in recent years [32]. It would be very interesting to expand this analysis to many other fields and extract the factors that determine their increasing or decreasing self-similar growth dynamics.

*

We thank Aric Hagberg for stimulating discussions. This work has been partially supported by the Office of Scientific and Technical Information (OSTI) of the U.S. Department of Energy. DIK was also supported in part by funds provided by the U.S. Department of Energy under cooperative research agreement DEFG02-05ER41360.

References

1. DEREK J. DE SOLLA PRICE, *Little Science, Big Science*, New York: Columbia University Press, 1963.
2. T. BRAUN, E. BUJDOSÓ, A. SCHUBERT, *The Literature of Analytical Chemistry: A Scientometric Evaluation*, Boca Raton, FL: CRC Press, 1987.
3. W. GOFFMAN, V. A. NEWILL, Generalization of epidemic theory: An application to the transmission of ideas, *Nature,* 204 (1964) 225–228;
   W. GOFFMAN, Mathematical approach to the spread of scientific ideas: The history of mast cell research, *Nature*, 212 (1966) 449–452;
   W. GOFFMAN, G. HARMON, Mathematical approach to the prediction of scientific discovery, *Nature*, 229 (1971) 103–104.
4. E. GARFIELD, The epidemiology of knowledge and the spread of scientific information, *Current Contents*, 35 (1980) 5–10.
5. A. N. TABAH, Literature dynamics: Studies of growth, diffusion, and epidemics, *Annual Review of Information Science and Technology (ASIS)*, 34 (1999) 249–286.
6. T. BRAUN, The epidemic spread of fullerene research, *Angew. Chem. Int. Ed. Engl.*, 31 (1992) 588–589.
7. M. E. J. NEWMAN, Scientific collaboration networks: I. Network construction and fundamental results, *Phys. Rev. E*, 64 (2001) 016131;
   M. E. J. NEWMAN, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E*, 64 (2001) 016132;
   M. E. J. NEWMAN, The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA*, 101 (2004) 404–409.
8. S. REDNER, Citations statistics from 110 years of Physical Review, *Physics Today*, 58 (2005) 49.
9. R. M. SHIFFRIN, K. BÖRNER, Mapping knowledge domains, *Proc. Natl. Acad. Sci. USA*, 98 (2001) 5183–5185.
10. C. CHEN, Searching for intellectual turning points: Progressive knowledge domain visualization, *Proc. Natl. Acad. Sci. USA* (suppl.), 101 (2004) 5303–5310.
11. K. W. BOYACK, R. KLAVANS, K. BÖRNER, Mapping the backbone of science, *Scientometrics*, 64 (2005) 351–374.
12. L. M. A. BETTENCOURT, A. CINTRON-ARIAS, D. I. KAISER, C. CASTILLO-CHÁVEZ, The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models, *Physica A*, 364 (2006) 513–536.
13. http://library.lanl.gov/lww/
14. F. BRAUER, C. CASTILLO-CHÁVEZ, *Mathematical Models in Population Biology and Epidemiology*. Texts in Applied Mathematics, 40, New York: Springer-Verlag, 2001.
15. K. L. COOKE, D. A. ALLERS, C. CASTILLO-CHÁVEZ, Mixing patterns in models of AIDS, In: O. ARINO, D. AXELROD, M. KIMMEL (Eds), *Mathematical Population Dynamics*, New York: Dekker, 1991, pp. 297–309;
   C. CASTILLO-CHÁVEZ, K. COOKE, W. HUANG, S. A. LEVIN, The role of long incubation periods in the dynamics of HIV/AIDS, part 1: Single population models, *J. Math. Biol.*, 27 (1989) 373–398.
16. A. H. GUTH, The inflationary universe: A possible solution to the horizon and flatness problems, *Phys. Rev. D*, 23 (1981) 347–356;
   A. D. LINDE, A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeniety, isotropy, and primordial monopole problems, *Phys Lett B*, 108 (1982) 389–393;
   A. ALBRECHT, P. J. STEINHARDT, Cosmology for Grand Unified Theories with radiatively induced symmetry breaking, *Phys. Rev. Lett.*, 48 (1982) 1220–1223.
17. A. H. GUTH, D. I. KAISER, Inflationary cosmology: Exploring the universe from the smallest to the largest scales, *Science*, 307 (2005) 884–890.
18. T. W. B. KIBBLE, Topology of cosmic domains and strings, *J. Phys. A*, 9 (1976) 1387–1398.
19. A. VILENKIN, E. P. S. SHELLARD, *Cosmic Strings and Other Topological Defects*, New York: Cambridge University Press, 1994.
20. S. B. PRUSINER, Scrapie prions, *Annu. Rev. Microbiol.*, 43 (1989) 345–374.

21. K. WEIGMANN, Fashion of the times, *EMBO Rep.*, 5 (11) (2004) 1028–1031.
22. S. B. PRUSINER, Prions, In: *Les Prix Nobel 1997*, Stockholm: Nobel Foundation, 1998, 262–323. Reprinted in *Proc. Natl. Acad. Sci. USA*, 95 (1998) 13363–13383.
23. http://www.who.int/csr/disease/avian_influenza/en/
24. S. IIJIMA, Helical microtubules of graphitic carbon, *Nature*, 354 (1991) 56–58.
25. T. BRAUN, A. SCHUBERT, S. ZSINDELY, Nanoscience and nanotechnology on the balance, *Scientometrics*, 38 (1997) 321–325.
26. T. BRAUN, S. ZSINDELY, I. DIÓSPATONYI, E. ZÁDOR, Gatekeeping patterns in nano-titled journals, *Scientometrics*, 70 (2007) 651–667.
27. R. P. FEYNMAN, Simulating Physics with Computers, *Int. J. Theor. Phys.*, 21 (1982) 467;
Quantum mechanical computers, *Found. Phys.*, 16 (1986) 507.
28. D. DEUTSCH, Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer, *Proc. R. Soc. Lond. A*, 400 (1985) 97.
29. P. W. SHOR, Polynomial time algorithms for prime factorization and discrete logarithms on a quantum computer, *Proc. 35th Ann. IEEE Symp. on Foundations of Computer Science*, S. GOLDWATER (Ed.), Los Alamitos, CA: Computer Society Press, 1994;
L. K. GROVER, A fast quantum mechanical algorithm for database search, *Proceedings, 28th Annual ACM Symposium on the Theory of Computing*, 212 (1996).
30. D. G. CORY, A. F. FAHMY, T. F. HAVEL, Ensemble quantum computing by NMR spectroscopy, *Proc Natl. Acad. Sci. USA*, 94 (1997) 1634–1639.
31. L. M. A. BETTENCOURT, J. LOBO, D. HELBING, C. KÜHNERT, G. B. WEST, Growth, innovation, scaling and the pace of life in cities, *Proc. Natl. Acad. Sci. USA*, 104 (2007) 7301–7306.
32. See, e.g., C. MODY, How probe microscopists became nanotechnologists, In: D. BAIRD, A. NORDMANN, J. SCHUMMER (Eds), *Discovering the Nanoscale*, Amsterdam: IOS Press, 2004, pp. 119–133;
L. ZUCKER, M. DARBY, Socio-economic impact of nanoscale science: Initial results and NanoBank, National Bureau of Economic Research, Cambridge, MA, Working Paper 11181 (2005); and the special issue of *Scientometrics*, 70 (2007) 541–880.

Appendix A
Bibliographical searches

Keyword and citation searches were performed using Search Plus, developed by the Los Alamos National Laboratory Reseach Library [13]. We found this to be superior in coverage, relevance, and accuracy than, e.g., Google Scholar, PubMed (for Prions and H5N1 Influenza), or SLAC-SPIRES (for physics).

In most cases, we constructed our databases from detailed keyword searches. However, for the subject of Cosmological Inflation, we relied upon citation searches rather than keyword searches in order to avoid overlap with the large volume of publications on economic inflation. The database for Cosmological Inflation was thus constructed from all publications that cited one or more of the following articles, which include landmark research articles and later review articles on the field:

- A. H. Guth, "Inflationary universe: A possible solution to the horizon and flatness problems," *Physical Review D* , 347–356 (1981)
- D. Linde, "A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy, and primordial monopole problems," *Physics Letters B* , 389–393 (1982)
- Albrecht and P. J. Steinhardt, "Cosmology for Grand Unified Theories with radiatively induced symmetry breaking," *Physical Review Letters* , 1220–1223 (1982)
- A. H. Guth and S.-Y. Pi, "Fluctuations in the New Inflationary Universe," *Physical Review Letters* , 1110–1113 (1982)
- J. M. Bardeen, P. J. Steinhardt, and M. S. Turner, "Spontaneous creation of almost scale-free density perturbations in an inflationary universe," *Physical Review D* , 679–693 (1983)
- D. Linde, "The inflationary universe," *Reports on Progress in Physics* , 925–986 (1984)
- D. La and P. J. Steinhardt, "Extended inflationary cosmology," *Physical Review Letters* , 376–378 (1989)
- K. A. Olive, "Inflation," *Physics Reports* , 307–403 (1990)
- V. F. Mukhanov, H. A. Feldman, and R. H. Brandenberger, "Theory of cosmological perturbations," *Physics Reports* , 203–333 (1992)
- D. Linde, "Hybrid inflation," *Physical Review D* , 748–754 (1994)
- L. Kofman, A. D. Linde, and A. A. Starobinsky, "Reheating after inflation," *Physical Review Letters* , 3195–3198 (1994)
- D. H. Lyth and A. Riotto, "Particle physics models of inflation and the cosmological density perturbation," *Physics Reports* , 1–146 (1999)

Resulting publications (both journal publications and conference proceedings) were stored in relational databases and checked for duplicates on entry.

For Cosmic Strings we used a query for search defined as (topolog* <in> Title/Subject/Abstract) <and> (Cosm* <in> Title/Subject/Abstract) <and> (string <or> defect <or> domain <or> texture <in> Title/Subject/Abstract). The results obtained were visually inspected for accuracy and several tests were performed for completeness by two of the authors who are domain experts in Cosmic Strings and Cosmological Inflation (LMAB, DIK). In particular, results from this search matched well with those of an earlier citation search, constructed akin to the one for Cosmological Inflation, based on early landmark papers and review articles on Cosmic Strings.

In publications for Prions and Scrapie we found, upon inspection, unanticipated records referring to birds of the genus Pachyptila, commonly also known as a "prion fairy." To remove this noise we formed the query string (prion <in> Title/Subject/Abstract) <and> (protein <or> amino <or> scrapie <in> Title/ Subject/Abstract) in relation to proteins or "amino," thereby eliminating the records which contains the information for birds. Records on scrapie were retrieved via the query string (<not> prion <in> Title/Subject/Abstract) <and> (scrapie <in> Title/Subject/Abstract) and merged the results of the above search, eliminating duplicates.

We used the query string (influenza <in> Title/Subject/Abstract) <and> (H5N1 <in> Title/Subject/Abstract) for H5N1 influenza.

In the case of Carbon Nanotubes, records were identified via the query string (<not> nanotubule <in> Title/Subject/Abstract) <and> (carbon <in> Title/ Subject/Abstract) <and> (nanotub* <in> Title/Subject/Abstract). This eliminated common references to unrelated cellular nanotubules.

Records for Quantum Computing and Computation were retrieved via the keyword search (Quantum Comput* <in> Title/Subject/Abstract).

Retrieved publications were parsed for author identification (including disambiguation), journal name, publication title, and year, and stored in relational databases.