Scientific discovery and topological transitions in collaboration networks

Citation: Bettencourt, Luis M.A., David I. Kaiser, and Jasleen Kaur. "Scientific discovery and topological transitions in collaboration networks." Journal of Informetrics 3.3 (2009): 210-221.

As Published: http://dx.doi.org/10.1016/j.joi.2009.03.001

Publisher: Elsevier

Persistent URL: http://hdl.handle.net/1721.1/50230

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Scientific discovery and topological transitions in collaboration networks

Luís M. A. Bettencourt[1,2], David I. Kaiser[3], and Jasleen Kaur[1]

[1]T-5 MS B284, Theoretical Division, Los Alamos National Laboratory
Los Alamos NM 87545

[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe Nm 87501

[3]Program in Science, Technology and Society, and Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 USA

Abstract:

We analyze the advent and development of eight scientific fields from their inception to maturity and map the evolution of their networks of collaboration over time, measured in terms of co-authorship of scientific papers. We show that as a field develops it undergoes a topological transition in its collaboration structure between a small disconnected graph to a much larger network where a giant connected component of collaboration appears. As a result, the number of edges and nodes in the largest component undergoes a transition between a small fraction of the total to a majority of all occurrences. These results relate to many qualitative observations of the evolution of technology and discussions of the "structure of scientific revolutions". We analyze this qualitative change in network topology in terms of several quantitative graph theoretical measures, such as density, diameter, and relative size of the network's largest component.

To analyze examples of scientific discovery we build databases of scientific publications based on keyword and citation searches, for eight fields, spanning experimental and theoretical science, across areas as diverse as physics, biomedical sciences, and materials science. Each of the databases is vetted by field experts and is the result of a bibliometric search constructed to maximize coverage, while minimizing the occurrence of spurious records. In this way we built databases of publications and authors for superstring theory, cosmic strings and other topological defects, cosmological inflation, carbon nanotubes, quantum computing and computation, prions and scrapie, and H5N1 influenza. We also build a database for a classical example of "pathological" science, namely cold fusion. All these fields also vary in size and in their temporal patterns of development, with some showing explosive growth from an original identifiable discovery (e.g. carbon nanotubes) while others are characterized by a slow process of development (e.g. quantum computers and computation).

We show that regardless of the detailed nature of their developmental paths, the process of scientific discovery and the rearrangement of the collaboration structure of emergent fields is characterized by a number of universal features, suggesting that the process of discovery and initial formation of a scientific field, characterized by the moments of discovery, invention and subsequent transition into "normal science" may be understood in general terms, as a process of cognitive and social unification out of many

initially separate efforts. Pathological fields, seemingly, never undergo this transition, despite hundreds of publications and the involvement of many authors.

Introduction

The general processes by which we collectively produce knowledge in science and technology have been the subject of endless fascination throughout history. Only in the last few decades, however, and especially after the electronic indexing of most publications in science and technology, have large-scale patterns in the development of science and technology [1-3] become available for quantitative analysis [2-4] and visualization [5-7]. This wealth of scientific and technical information is huge, messy, and often characterized by significant gaps and inconsistencies. Nevertheless, its existence and development beckons us to start testing classical ideas from the history, philosophy, and sociology of science, as well as develop new ones, capable of shedding light on the global enterprise of science and technology. We can reasonably expect that answers to these questions would have immense impact on fundamental aspects of social and cognitive sciences, as well as suggest general mechanisms for the origins of economic and technological growth and point to public policy that could accelerate discovery and help us more promptly reap the benefits arising from scientific development.

We clearly appreciate, at least since the work of Thomas Kuhn [1] but also from our personal experience, that the practice of science and technology is not conducted by fully rational agents, but rather by individuals driven by their personal passions and limited views of the world and the fields they work in. Despite these facts, and the difficulty to model behaviors that are not in some sense optimal, it is now clear that there are many important regularities in bibliometric data [2-4]. For example de Solla Price [2], one of the founders of bibliometrics, observed very early on that scientific fields tend to grow exponentially in their early phases before reaching a period of saturation. Science as a whole has also grown roughly exponentially, at a faster rate than the human population, implying that most scientists throughout history are active now, and that paradigms of the past are re-learned, re-examined and put to work by an ever growing numbers of new scientists. The recency of the bulk of the world's scientific population creates curious bibliometric effects, such as skewing the statistics of citations toward recent papers.

Aiming to discover the mechanisms whereby scientific ideas are created and spread, Goffman, Garfield and others suggested long ago [8-11] that a fruitful approach to understanding the temporal development of scientific fields would be to model them mathematically as populations exposed to a contagious agent, implying that scientific ideas and concepts may behave much like an infectious "pathogen" in a population of susceptible individuals. We have recently tested this hypothesis [12,13] to show that mathematical models of epidemiology – suitably adapted to describe some of the social dynamics of science – do indeed give an extraordinarily good description of developing scientific fields, albeit with parameters typical of very slow, hard to catch, communicable diseases.

The existence of new and growing amounts of bibliometric data is inspiring a variety of other approaches to identify global interdependencies between fields [5,6], detect temporal patterns of change [14-16] and to model network structures of collaboration, and citation [17-20]. These studies begin to suggest that the advent of a new field, resulting from conceptual or technical discovery, may be identifiable early on. If possible, such early identification would allow scientific communities, funding program managers, and policy makers to identify promising yet fledgling areas of new research, and to spur on new discoveries.

The issue of identifying the emergence of new fields is simultaneously very interesting and extremely difficult. The main issue deals with the early identification of a new research theme, expressed in terms of a few publications, against a gigantic background of thousands or millions of others. Early on, the jargon that may later accompany new fields is typically not yet settled and often refers to older concepts, making identification based on natural language analysis very difficult, at least with current technology in computational linguistics. Moreover, most new research directions are red herrings, which raises the difficult issue of how to identify new revolutionary scientific and technological ideas out of large number of dead ends.

Several approaches have now been proposed to detect novelty and/or the advent of a new field. Motivated by the analysis of temporal trends on the world wide web, Kleinberg [14] suggested that temporal bursts of activity in documents sharing certain key terms (say in publications on a theme) may signal a new discovery. There are however many confounding effects. Many such bursts are now known to be the result of hype, resulting from postings in blogs, message boards, and other popular venues, which do not necessarily reflect the judgment of the (usually smaller) informed communities of scientific practitioners. Another approach relies on the analysis of citation lists, and is based on the idea that a field may be identifiable by reference to a group of influential papers. Chen [15] explored the idea that fields (including emerging ones) may be characterized by patterns of co-citation of several founding papers, and constructed a suggestive visualization of this effect for the field of botulinum toxin.

The present paper is dedicated to a complementary approach, which relies primarily on measuring the changes in the structure of collaboration (and, implicitly, of temporal growth) of an emerging field. The literature on the history and sociology of science has amply demonstrated [21-25] that new conceptual or technical breakthroughs typically lead not only to rapid growth in the number of scientists and publications in a field but also to tighter collaboration, made possible through the new set of shared concepts and techniques. This change in the network of collaboration, it turns out, is a measurable and general effect, and its quantitative characterization is the subject of this paper. To proceed we adopt a manifestly empirical approach based on the bibliometric analysis of the evolution of eight fields in science and technology, across physics, biology, and material science.

Before we do so, however, we must take a short detour to discuss what is known about the processes whereby new fields arise. At the qualitative level the process was described most famously by Thomas Kuhn in his seminal work, *The Structure of Scientific Revolutions* [1]. Kuhn analyzed several historical examples of discovery, especially from the history of physics and chemistry. He contended that new (and eventually) fruitful fields arise from two special and closely related processes: *discovery* and *invention,* see Figure 1. Discoveries are often fortuitous, and involve small numbers of scientists. They may initially not reveal their full conceptual and experimental implications. Kuhn used the example of the discovery of oxygen in the 1770s, probably independently by C. W. Scheele in Sweden, Joseph Priestley in Great Britain, and Antoine Lavoisier in France. In this initial period, fields are small and efforts are independent; there are typically incommensurate interpretations of the discovery. Its ultimate theoretical explanation and practical applications remain largely unclear or may even be plain wrong (as judged from hindsight). During the process of *invention*, the explanatory or practical potential of the idea becomes clear; this stage is associated with the beginning of large-scale adoption and ensuing widespread collaboration. This is typically the time when a new theoretical framework or technological design emerges (what Kuhn called a *paradigm*), which allows a large community of researchers to share a common language, collaborate, exchange junior scientists, and so on. The process of invention in the case of oxygen stemmed from Lavoisier's quantitative experiments, which elucidated the chemical nature of combustion as oxidation. This conceptual advance opened the doors to the understanding of the fundamental role of oxygen in many common reactions, from inorganic to organic chemistry, including respiration and metabolism.

The process of discovery ushers in what Kuhn (in)famously labeled *normal science*, the pursuit of research within a broadly defined conceptual paradigm. In Kuhn's account, a reigning paradigm will hold within a scientific community until a critical mass of anomalies accumulates. These anomalies, unexplainable within the existing paradigm, will precipitate a crisis. The period of crisis will end, in Kuhn's view, as abruptly as it began, with a scientific revolution that replaces the old paradigm with a new one, wholecloth. And then the cycle will repeat itself, with normal science chugging along under the new paradigm until it, too, succumbs to a period of crisis, revolution, and replacement.

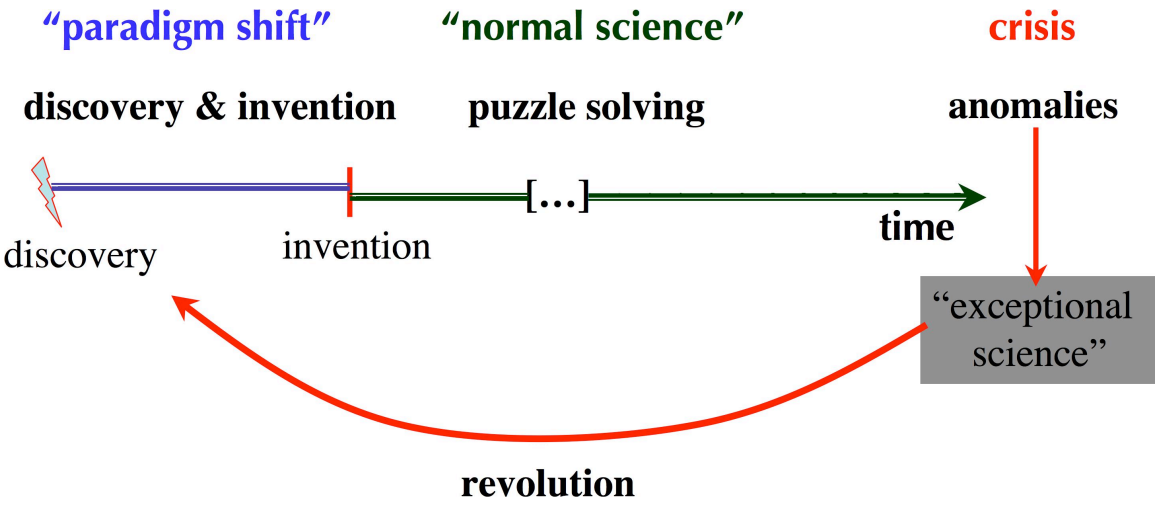

Figure 1: A schematic representation of Kuhn's account of the advent, development, and dissolution of scientific fields. The scheme and nomenclature are from [1].

Kuhn's work is now nearly fifty years old; its influence on the history, philosophy, and sociology of science has been profound. Like any provocative work, *The Structure of Scientific Revolutions* has attracted sustained criticism, most details of which need not concern us here. Most important have been critiques of Kuhn's central notion of paradigms, which – as even Kuhn admitted in later years – smuggled in a series of notions that might have been better kept distinct. [26] Did "paradigm" refer primarily to an overarching conceptual framework or to a body of shared techniques and practices? [25,27] Were paradigms truly homogenous systems of belief or practice, shared equally by all practitioners of a given science at a given time, or does significant substructure exist within any scientific community? [24] Were concepts within competing paradigms genuinely incommensurable – literally untranslateable into each other – and hence doomed to remain forever closed off from one another, despite continuities in experimental design or instrumental practice? [24,28]

Overlooked in these otherwise cogent commentaries and critiques, however, lay a valuable insight we may still borrow from Kuhn. For Kuhn had suggested, even before large-scale bibliometric measurements were possible, that general dynamical and social processes underpin the creation of scientific and technological knowledge, regardless of disciplinary specificity. One need hardly remain wedded to every feature of Kuhn's account to appreciate this perceptive proposal, and seek to build upon it.

The central thesis of this paper is that the creation and spread of new discoveries through a scientific community creates qualitative, measurable changes in its social structure. Bibliometric studies have started to buttress this view quantitatively. These changes have been noted anecdotally in the past both in science, for example in the community of RNAi researchers [29], and in other creative fields, such as the structure of collaborations that underpin the making of Broadway musicals [30]. Here, as in these other examples, we isolate several global properties of networks downstream from a discovery, which remain common across diverse fields. In particular, networks of researchers manifest a

topological transition from an initial number of small clusters of scientists to a giant component of collaboration, see Figure 2. This suggests that as scientific and technological discovery takes root among a community of practitioners, it leads to a large-scale reorganization of the social structure of collaboration, akin to a universal critical phenomenon in physical systems (albeit in a finite system).

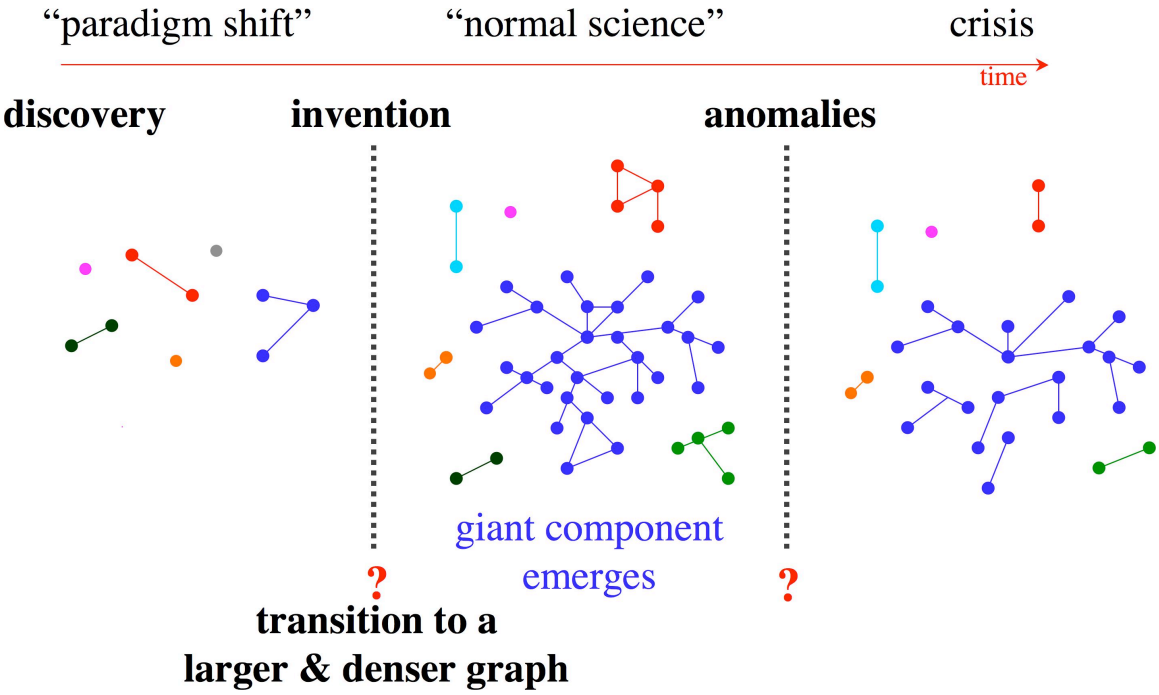

Figure 2: New scientific concepts lead to temporal and structural rearrangements in the structure of scientific collaborations. Successful fields of science typically start as small independent efforts (shown as connected components of different colors), which grow (differentially) and eventually form a giant component in a collaboration graph. Although the size of a field and its rate of growth vary, this structural transition is hypothesized to be universal.

The remainder of this paper is organized as follows. In the next section we give the details of our bibliometric data, its collection, parsing and organization. We also clarify how author names are identified and how network structures of collaboration are measured. We then show the temporal evolution of these quantities for eight distinct fields in science and technology, pointing to some of the similarities and differences among them. Finally we discus our results and their implications for fleshing out a new quantitative "science of science." We also point to future work, necessary to evaluate the hypotheses advanced here more thoroughly.

Data and Methods

We have assembled bibliometric data for eight fields, one of them usually considered pathological, namely cold fusion. We have retrieved a collection of publications (both journal articles and conference proceedings) using Los Alamos National Laboratory's

SearchPlus, which aggregates information from the following databases: BIOSIS®, Engineering Index®, Inspec®, ISI Proceedings®, ISI SciSearch®, ISI Social SciSearch® and ISI Arts & Humanities®. We have also explored using Google Scholar but found a larger number of misclassified records and omissions, which led us to forego it as a source of bibliometric information for scientific studies at the present time. We do not have access to Scopus through our research library, and were thus unable to compare its results to those aggregated by SearchPlus.

Records are retrieved during May 2008 in XML and parsed and archived in relational databases. Articles are indexed by title, publication identifier (SearchPlus), journal or conference publication information, and year. A separate table with author information and corresponding affiliation (whenever available) is also constructed and linked to each publication. We parsed names of authors following standard conventions of keeping only first initial and last name. While this choice may associate together separate authors, it reduces issues of misclassification of many possible spellings and middle names of the same author as separate identities. A summary table of publication numbers by fields is given in Table 1. Select temporal evolution of numbers of authors and publications by field is shown in Figure 1, more details are given in [13].

Field	Years	Number of Publications	Number of Authors
Cosmological Inflation	1981-2005	5135	3410
Cosmic Strings	1976-2005	2443	2292
String Theory	1974-2005	9766	25022
Carbon Nanotubes	1992-2005	30521	25464
Quantum computing	1967-2005	8946	7518
Scrapie and Prions	1960-2005	11074	14620
H5N1 Influenza	1984-2005	604	1281
Cold Fusion	1980-2005	871	1637

Table 1. Summary statistics for eight scientific and technological fields growing from initial conceptual or technological breakthroughs.

As has been amply described elsewhere, publications and their authors form a bipartite graph [15-20,33-35]. To construct networks of co-authorship we simply established a binary link between any two authors that ever published together in the same field, effectively projecting the bipartite graph of publications and authors over the space of authors. The resulting binary graph is easy to analyze by standard techniques. In the next section we give results for the evolution of each field in terms of the number of edges vs. number of nodes, diameter (the length, in number of edges, of the longest geodesic path between any two vertices.) and fraction of nodes and edges in the largest component of its collaboration graph. These results were obtained via the standard application of NetworkX (https://networkx.lanl.gov) to the graphs for each field constructed from its database.

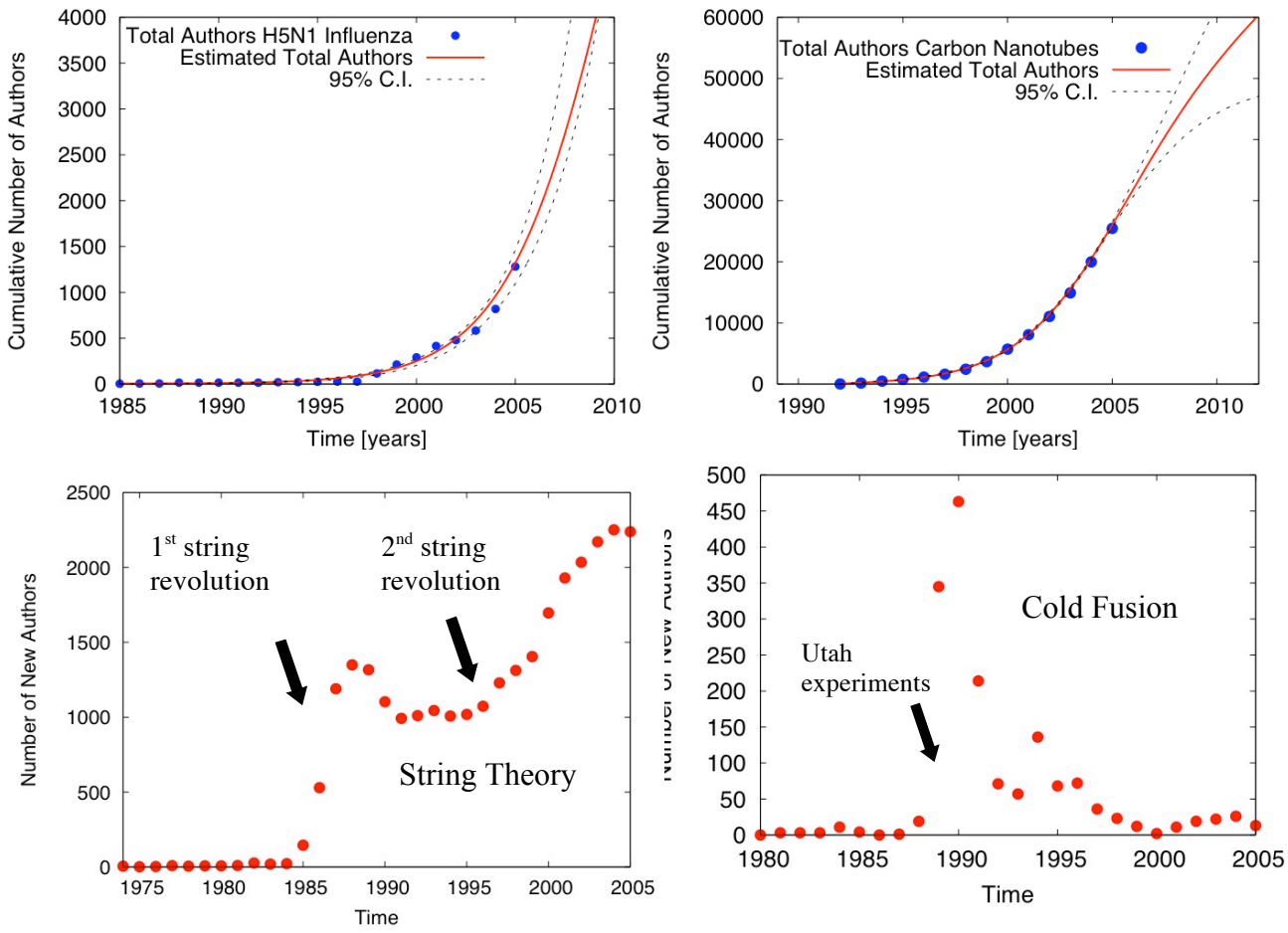

Figure 3: Example time series for number of authors in four fields. The cumulative number of authors in H5N1 influenza (top left) and carbon nanotubes (top right) grew quickly after the first cases in humans and the invention of a technique to grow fullerenes in a particular shape, respectively. The temporal growth of these and other fields is very well described by population models, suitably adapted from epidemiology (solid and dotted lines, see [13]). Lower panel shows the number of new authors per year for superstrings (bottom left) and cold fusion (bottom right), where we also signal moments where, with hindsight, fields went through new conceptual transformations (string theory) and experimental claims of a new discovery (cold fusion).

Results

Several global properties of emerging fields, expressed in terms of their collaboration networks, show non-trivial behavior. First, collaboration graphs "densify," that is, show an increase in the average number of edges per node. Interestingly, this occurs empirically as a self-similar process well fit by a power-law scaling relation.

Simultaneously, graph diameters grow and eventually stabilize, and the network as a whole undergoes a topological transition in which a giant component emerges once average connectivity reaches a critical threshold, which is field specific.

Densification and growth

It is a remarkable property that when fields grow, their networks of collaboration also become denser. This means in practice that the average number of edges per node tends to increase over time. Empirical analysis of this effect [31] reveals that the relation between number of edges and nodes is often self-similar and can be described in terms of a simple scaling law

$$\text{edges} = A\ (\text{nodes})^{\alpha}, \qquad (1)$$

where A and α are constants. The scaling exponent, α, expresses the densification effect in a way that is independent of scale (number of nodes). This can be shown explicitly by considering the number of edges in a collaboration graph at two scales (in terms of the number of the number of nodes), N and λ N:

$$\text{edges}(\lambda\ N) / \text{edges}(N) = \lambda^{\alpha}, \qquad (2)$$

so that $R = \lambda^{\alpha-1} - 1$ is the corresponding fractional *increase* in the number of edges (assuming $\alpha > 1$, see below) .

We find that most fields show a percent increase R, on the order of 9-30%, as their size doubles, with fields in physics (cosmological inflation, string theory, comic strings) showing greater densification, followed by quantum computation and carbon nanotubes, and finally the biomedical fields of scrapie and prions and H5N1 influenza. Exponents are summarized in Table 2; a few example plots are given in Figure 4. In this sense the scaling law for densification, while varying quantitatively in terms of the value of the exponent, α, nonetheless followed a general form across the various fields.

It is noteworthy that research areas that do not possess a high degree of shared concepts or practices tend to densify more slowly, if at all. Several examples among our set of fields illustrate this effect (see also Figure 4). For many years, prior to 1994-95, quantum computing was a field without a unified conceptual framework, and as a result even as it grew slightly it did not densify. Similarly H5N1 influenza is a new field created to a large extent by the increasing number of human infections, their high mortality rate, and the perceived threat of a looming high-mortality influenza pandemic akin to that of 1918. In this sense the field is being driven by societal concerns and may not yet have created new concepts or experimental techniques that distinguish it sufficiently from research in neighboring areas, such as on other influenza strains. It densifies very slowly with an exponent barely above unity. Finally, nuclear cold fusion is a field that never found a solid experimental or conceptual proof of principle, and as such has never become a field of collaboration and exchange. It shows $\alpha = 1$, manifesting the fact that it is mostly the product of small, disparate, and often incommensurate efforts.

Thus densification ($\alpha>1$) in the aftermath of new conceptual or technical practices seems to be a necessary, but possibly not a sufficient condition (see below) for a successful field to form and progress into the stage of normal science.

We close this section by discussing the consequences of increasing local connectivity per node for the global structure of a graph. The phenomenon bears some analogy with percolation theory. On a lattice with a fixed number of spatial neighbors, increasing the probability of being connected to each one of them (the average degree) eventually leads to a phase transition, in which an infinite connected cluster or component emerges. However, several caveats must be considered in order for a strict analogy with percolation to hold. First, collaboration graphs are typically heterogeneous, with degree distributions that show power law tails [15-20, 33-35]. Second, while the number of edges per node increases for successful fields, the graph density (the ratio of the number of actual edges to all possible edges in a graph with the same number of nodes) invariably decreases. This is because in a binary undirected graph the number of possible edges between N nodes grows as $N(N-1)/2 \sim N^2$, while, as we have seen above, the densification scaling exponent is $1<\alpha<2$.

Under these circumstances, can fields indeed become connected, in the sense of developing a giant component of collaboration? How do global properties of collaboration graphs behave as fields grow? Below we describe what actually happens and investigate whether there is universal behavior in general for emerging scientific fields, independent of the field under consideration.

Cosmological Inflation	$\alpha=1.38 \pm 0.02$
Cosmic Strings	$\alpha=1.21 \pm 0.01$
String Theory	$\alpha=1.36 \pm 0.01$
Carbon Nanotubes	$\alpha=1.17 \pm 0.01$
Quantum Computing	$\alpha=1.21 \pm 0.01$
Scrapie & Prions	$\alpha=1.12 \pm 0.01$
H5N1 Influenza	$\alpha=1.05 \pm 0.03$
Cold Fusion	$\alpha=0.99 \pm 0.02$

Table 2: Densification exponents for eight scientific fields. All fields with conceptual or experimental frameworks grow and densify (i.e, show $\alpha>1$), whereas fields in search of breakthroughs do not ($\alpha \sim 1$), such as cold fusion. Parameter estimates were obtained via ordinary least squares regression to a linear relation in a double logarithmic plot.

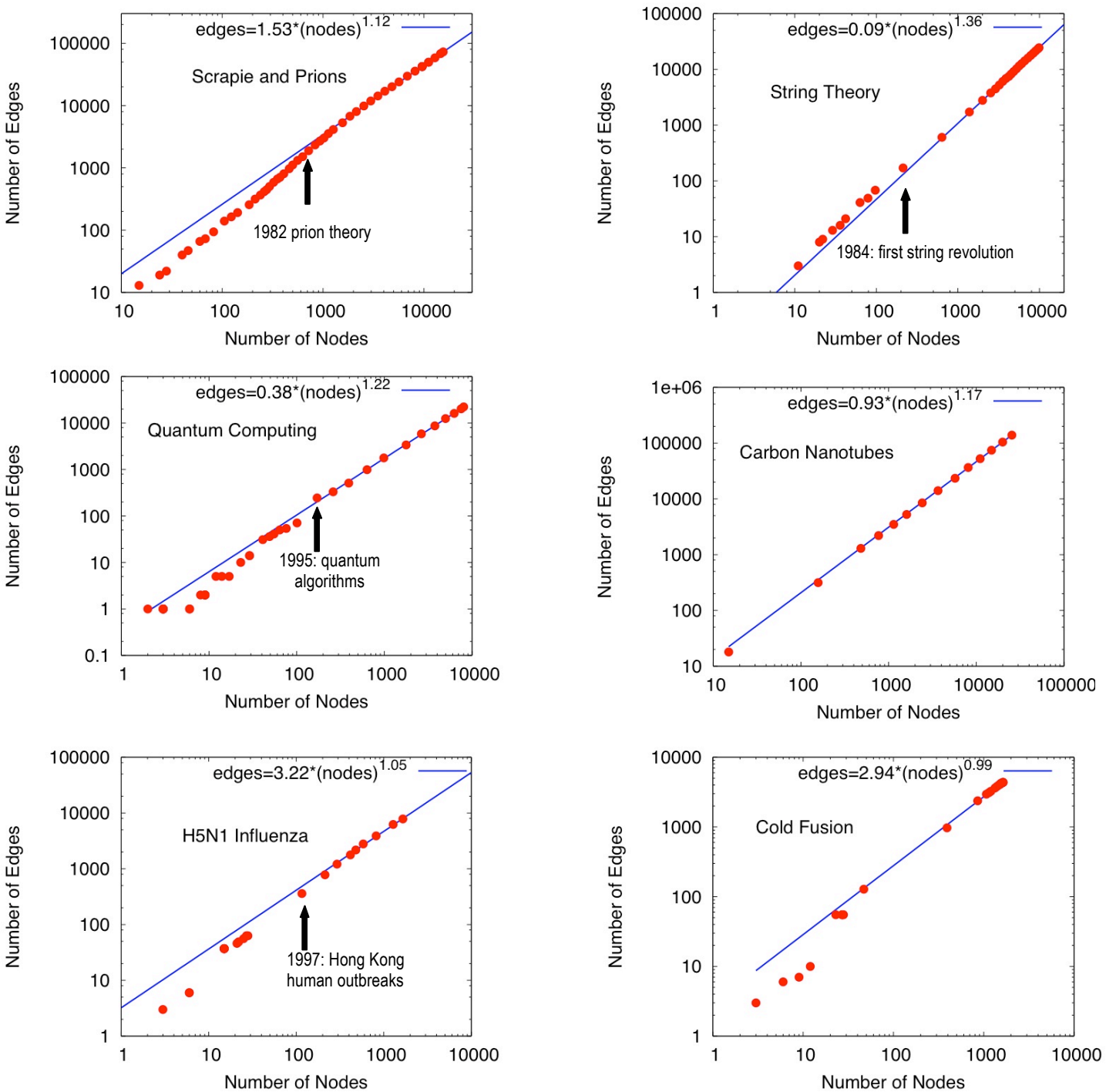

Figure 4: Densification of collaboration graphs (increasing number of edges per node) for six fields. All fields with a robust set of shared concepts and techniques show a scaling exponent ($\alpha>1$). Fields motivated by common goals (cold fusion) or driven primarily by societal needs (H5N1 influenza) do not show significant increase of the number of edges per node as the field grows.

Network diameter

Another interesting quantity that measures global properties is the collaboration's graph diameter. In more general circumstance (such as all patents registered with the United States Patent and Trademark Office, or citation and affiliation graphs for all preprints collected in arXiv.org), Leskovec, Kleinberg & Faloutsos [31] found that the network diameter tends to decrease as a graph grows. The diameter of a graph is the average path length (measured in number of edges) between two nodes, so that its decrease implies a greater number of neighbors within the same number of edges, and thus a more tightly woven community of collaboration.

A few simple examples of how the diameter of a graph d changes with the number of nodes may illustrate these points. For a simple linear chain with N nodes the graph diameter would grow linearly with N, $d \sim N$. On the other extreme, a star graph would have constant diameter $d \sim 2$, regardless of N. In small-world graphs typical diameters scale logarithmically, or more slowly, with the number of nodes, N. Several classes of scale free networks have diameters that scale like $d \sim log\ (log\ N)$ [36] and, in these circumstances, are known as ultra-small graphs.

Collaboration graphs for our scientific and technological fields do not behave in any of these relatively simple ways. In fact most fields show an initial fast growth in their diameter, which then tends to stabilize and stay approximately constant $d \sim 12\text{-}14$. We have verified that the diameter does not continue to increase as the logarithm of N, as one might expect in a small-world graph. It is more difficult to exclude that the diameter does not increase on average more slowly, like $log\ (log\ N)$, though clearly this does not fit well any of the specific patterns of Fig. 5. Thus, even as the collaboration graph densifies, the graph stays globally connected such that the diameter of its largest component does not change measurably. What are the consequences of these properties for the global structure of collaborations?

Discovery and topological transitions in collaboration networks

A collaboration graph that densifies with constant or decreasing diameters suggests that a global topological transition may occur in the graph as a whole as it grows. Topological transitions of this nature have been studied in general terms, usually by analogy to percolation phase transitions in statistical physics [37,38]. While the parallel is imperfect in some respects, the thorough understanding of these phenomena in statistical physics allows us to make informed hypotheses about the phenomenology of real networks undergoing topological transitions.

In statistical physics bond percolation phenomena consider a number of sites (analogous to nodes), which may or may not be connected (via an edge or bond) to their neighbors with a probability p. As p increases from zero (depending on the dimensionality and geometry of the lattice of sites) a critical value $p=p_c$ will generally be reached, at which a spanning connected cluster of sites emerges. This spanning cluster is analogous to a giant component in a graph. In critical phenomena the thermodynamic limit, when the size of

the system goes to infinity at fixed p, plays a fundamental role in establishing the existence and type of a phase transition. Only in this limit can a truly infinite spanning cluster be formed, and scaling analysis at different system sizes usually allows the extrapolation to this limit and the determination of p_c, and of associated critical exponents. These numbers characterize how the relative size of the spanning cluster P, and fluctuations around it, change with p in the vicinity of p_c. In percolation theory, in the vicinity of the critical point ($p>p_c$), P follows a relation like

$$P(p) = P_0 (p-p_c)^{\gamma} , \quad (3)$$

where P_0 is a constant and γ is the scaling exponent. Note that in general P_0 and p_c depend on details, while γ may be common to many models. Whenever this happens one speaks of universality, in the sense that models that are different in detail exhibit the same form of critical behavior [38].

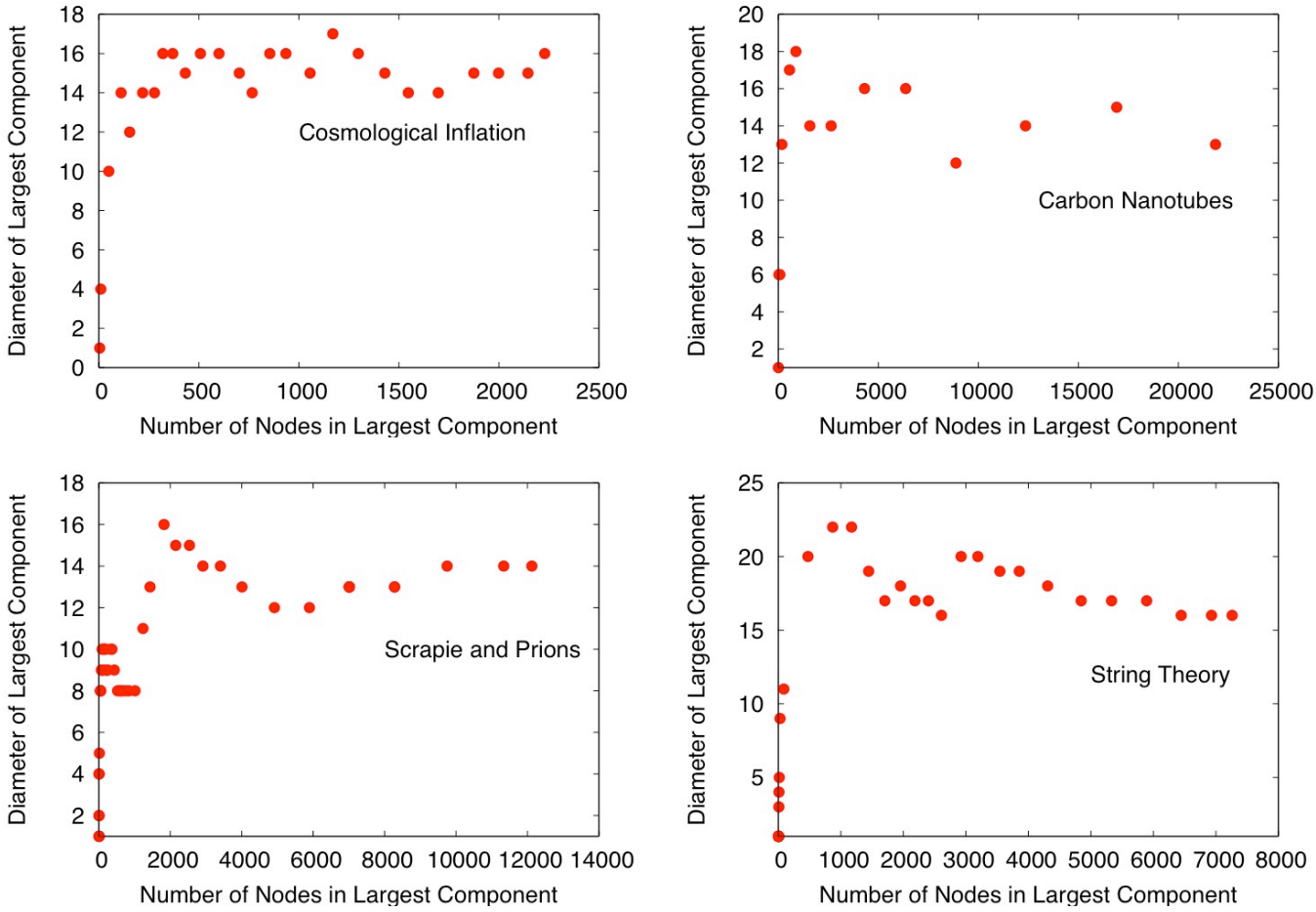

Figure 5: Change in network diameter with graph size (measured in number of nodes) for several growing scientific fields. Network diameter tends to grow fast as the field first forms, but eventually stabilizes to an approximately constant value.

Here we want to observe, and give some empirical evidence for, the analogy between percolation phenomena and the formation of giant components of collaboration in the

aftermath of a discovery. In the context of scientific discovery, unlike that of purely statistical models of percolation, the finiteness of the graph, as well as its evolution in time, are necessary characteristics of the phenomenon and make its analysis more challenging. Nevertheless, as we show below, there are clear signs that universal behavior common to all scientific and technological fields may be at play, characterizing scientific discovery as a universal critical phenomenon of the same kind in their social network of collaboration.

To motivate this discussion we first show, in Figure 6, the temporal evolution of the fraction of edges in the largest connected component. This is the natural analogue of the size of the percolating cluster in statistical physics and has been the proposed order parameter in several studies of percolation in networks [33]. The fraction of nodes in the largest component behaves similarly. The examples of Figure 6 show behavior typical of every successful field that comes to establish central conceptual or experimental techniques and grow. Initially the largest component is small and the fraction of edges that belong to it can be large in relative terms. However, as the field starts to grow and alternative approaches are explored, several small clusters of collaboration form, resulting in a small value of the order parameter. The process of invention, in which a new set of concepts and techniques unifies different approaches and enables more widespread collaboration, marks the beginning of the emergence of a giant connected component, which then continues to grow in relative terms as the field matures. That the advent of new concepts (and in some cases other events, such as new public emergencies) can restructure collaboration networks is clear from Figure 6, which show how string theory and the concept of prions lead to more interconnected collaboration networks, even in already established fields.

Finally, we show in Figure 7 that all fields undergo a similar topological transition, characterized by an exponent $\gamma \sim 0.35$, but with different values of P_0 and p_c. This can be seen by collapsing all curves for the dependence of the size of the largest collaboration cluster vs. the difference between average node degree, $\langle k \rangle$, to its critical value, k_c, which assumes the role of p_c and varies from one field to another. This suggests that all topological transitions may belong to a common universality class, but that their small scale structure, manifest e.g. in the value of k_c, is distinct. This suggests that the general behavior manifest in this topological transition is an emergent large scale phenomenon, but that it does not provide per se a model for the detailed dynamics of each field. These interesting questions will be analyzed in a more technical forthcoming publication. Fields still in search of a unifying set of concepts or techniques do not show a transition, as we also show in Figure 7.

Discussion and conclusions

It has long been a goal of the history, philosophy, and sociology of science and, more recently, of bibliometrics and a new "science of science," to identify (quantitative) indicators or circumstances that reveal moments of scientific and technological discovery. Although discovery involves complex social and cognitive processes and may not be predictable in detail, a better understanding of the circumstances that lead to rapid changes in the conceptual and technological makeup of knowledge may reveal general

processes – at both the individual and group levels – at play across fields. A greater understanding of these dynamics may make possible strategies, undertaken by scientific communities, scientific funding agencies, and policy makers, to accelerate the processes of scrutiny, analysis, and development associated with transformative new concepts. Thus, a more quantitative "science of science" may allow society to reap the benefits of new discoveries sooner, and encourage the processes whereby it takes place, especially in areas of critical need such as health, energy, climate change, and so on.

Clearly the advent of novel concepts and techniques creates new opportunities for scientists and inventors by increasing the scope for the creation and transmission of knowledge. These changes impact temporal patterns of publication [12,14] and citation [15], and introduce jargon [39-40], thereby affecting, as we have shown here, the manner in which scientists collaborate with each other. With an increasing amount of bibliometric data at our disposal, it is therefore possible that these changes – already discussed in the classical literature of the history, philosophy, and sociology of science – can be measured. Taken individually or together, they can help pinpoint quantitatively moments of discovery and invention, as well as their preceding and subsequent dynamics.

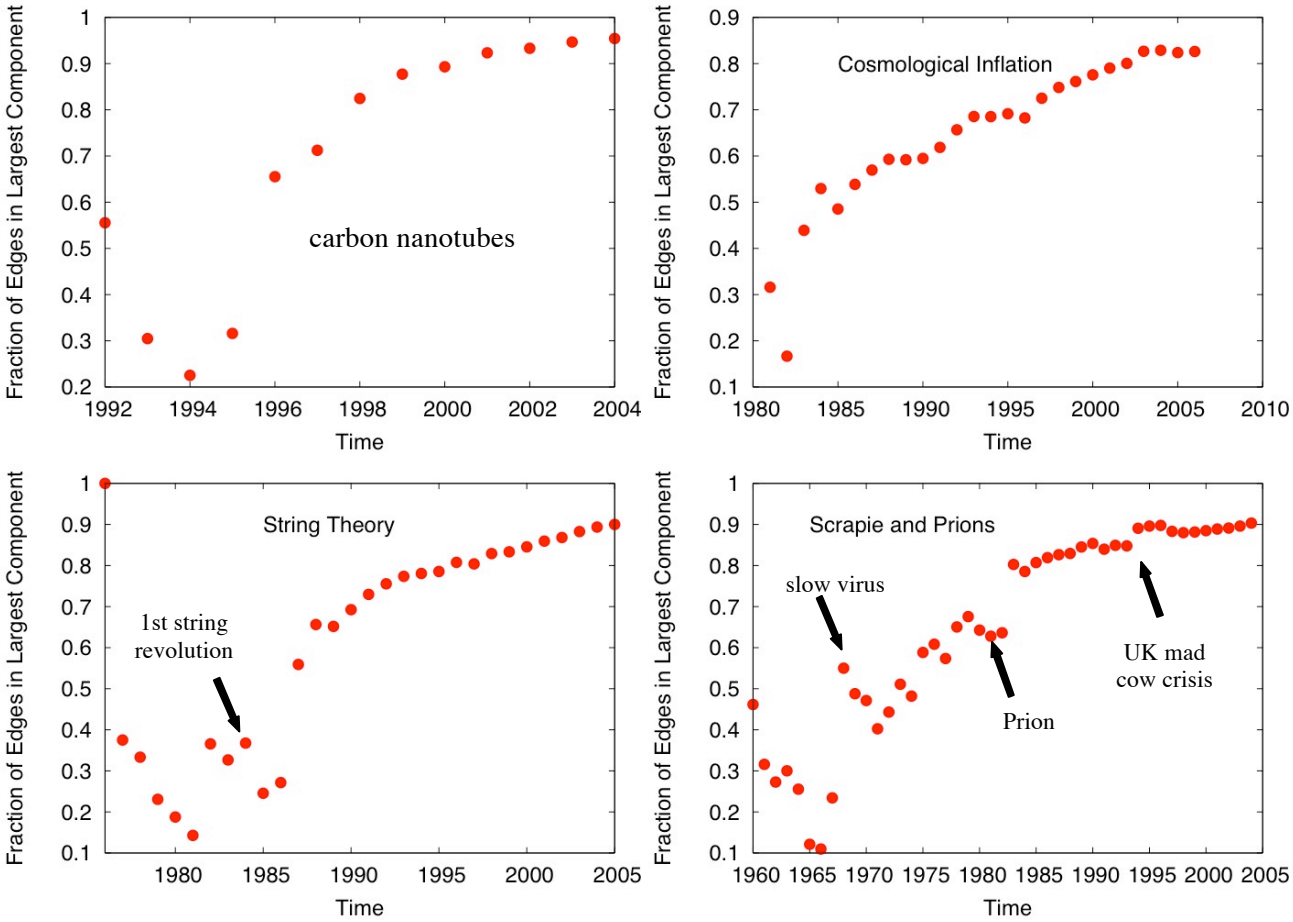

Figure 6: Time series for the fraction of edges in the largest component. The increase in the fraction of edges in the largest component of emerging fields suggests that the introduction of new concepts and techniques leads to a topological transition where most scientists in the field become connected by ties of collaboration**.**

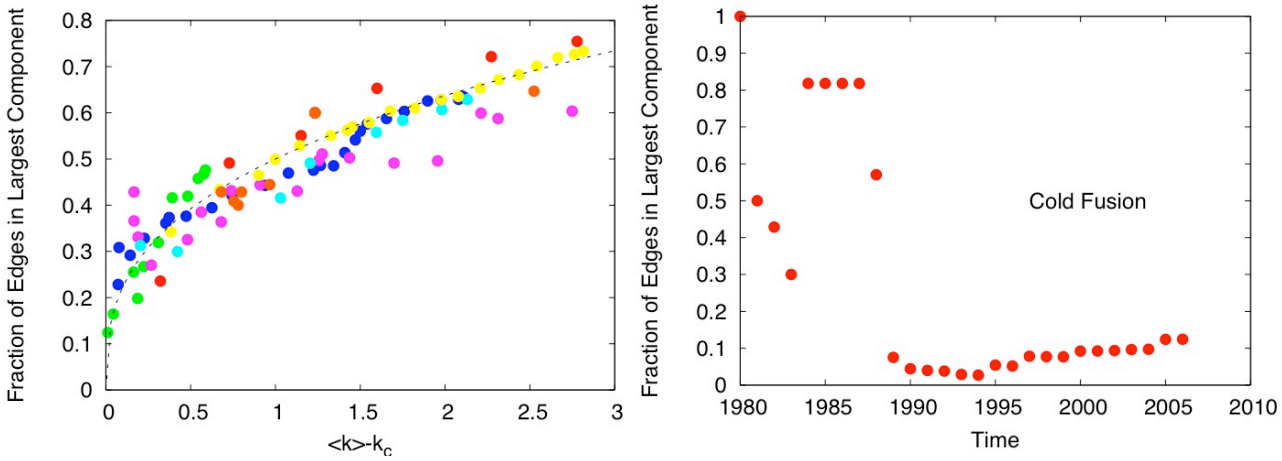

Figure 7: Field development and topological critical behavior. (left) All successful fields (different colors) display the same approximate critical behavior, described by Eq. (3) (dashed line γ=0.35), for the relative size of the largest component, in terms of $p=\langle k\rangle$, taken as the average degree. The critical k_c is not universal and varies from one field to another. Fields without an established (and shared) set of concepts and techniques, such as cold fusion (right), do not display a topological transition.

Each of these measures of discovery suffers from some practical shortcomings. The moment of discovery is typically messy, in the sense that the value of the new concept or technique often takes time to crystallize and be appreciated by competitors, permeate the language, collect citations, and spur on new collaborations. In most historical studies of scientific discovery it is only with hindsight that a great discovery is hailed as such, and it is often years before it is recognized as the new centerpiece for an entire scientific field [41,42]. Moreover, from a purely practical point of view, early publications in a new field are difficult to identify as such by automatic means, because their language, references, and so on often pool information from several existing efforts and are thus far from unique or novel. In this sense any search for very early publications in a field, without reference to its subsequent history, will be contaminated by noise, and is in our present opinion probably impossible. It is for these reasons that we went to some length to distinguish the process of *discovery* from that of *invention* (as defined by Kuhn), where indeed the new developments become useful and are quickly adopted. It is in our opinion invention that can be better measured quantitatively. Discovery may in turn be traced back from invention, usually without much difficulty, given the typically small size of any field during its initial stages.

This paper suggests that there may be a viable alternative to attempting to judge if the contents of publications are new and seminal, or indeed to identifying new fields via patterns of co-citation. The idea is that new fields nucleate around unifying concepts and techniques that allow them to both grow and exist as a community of shared concepts and practices. Because of these shared concepts and practices, collaboration becomes more widespread and leads to the emergence of a giant component in a graph of co-authorship. We emphasize that if more informal measures of scientific exchange could be measured – such as discussions or other personal exchanges – we would expect a qualitatively similar result.

The strength of identifying a critical phenomenon in collaboration as the signal for discovery and invention is that it is both robust and general. We have indeed shown here that nearly all fields under study show the same qualitative behavior as they evolve, and that once properly scaled, their topological transition is characterized by the same scaling of the fraction of edges in the largest component (the order parameter) in the vicinity of the critical point associated with increased connectivity degree. Note that each field is characterized by different values of p_c and P_0, as well as different sizes, clustering, and other network properties. This is similar to what happens in statistical physics, in which different microscopic models display phase transitions at different values of temperature, pressure and so on, but whose critical scaling behavior falls within the same "universality" class.

We may posit a plausible explanation for this universal behavior. During the second half of the twentieth century, scientific and technical education around the world came to rely more and more heavily upon a phase of postdoctoral training. To the scientists and policymakers who advocated for and designed the postdoctoral stage, the main point was to generate *circulation* of young scientists between various schools and groups. Backed by powerful institutions in the United States and Western Europe, the postdoctoral system quickly spread to become the norm worldwide [25]. With a growing fraction of all scientific authorship undertaken by postdocs, and with postdocs moving every few years to new institutions, postdocs in effect knit together distant authors and institutions; little wonder that we should see significant overlaps in co-authorship among practitioners of most scientific and technical fields. This general circulatory-co-authorship mechanism, common across virtually all fields of science and technology today, helps explain (or at least motivate) the universality-class phenomenon emerging from our co-authorship network analyses, especially since all of our examples come from the last decades of the twentieth century, well downstream from the major institutional innovation of introducing and standardizing postdoctoral training across fields and around the world.

This same shared feature – widespread circulation of postdocs – may also help explain the hierarchy of densification exponents, α, among the various fields surveyed here. The three areas of theoretical physics that we investigated (cosmological inflation, cosmic strings, and superstring theory) all showed values of α significantly greater – by as much as one-quarter to one-third – than the values of α for the biomedical fields (scrapie and prions, and H5N1 influenza research). This range is well correlated with the increased duration per postdoctoral appointment in biomedical fields, which have grown to be roughly one-third longer, on average, than those in physics and astronomy [43]. This leads to a proportionally smaller rate of circulation during any fixed time period for young researchers in biomedicine than in physics, and hence presumably a smaller rate of co-authorship with people at new institutions. While obviously preliminary and in need of further analysis, this hypothesis might account for both the universal structure and differential rates of densification in co-authorship networks across the recent sciences.

Our suggestion here that there may be a universal character to discovery and invention is clearly only based on the circumstantial evidence from the eight fields studied. However other analysis of developing fields such as RNAi [29], and even Broadway musical collaborations [30] point, at least qualitatively, in the same direction. On the other hand, the evidence reported in [31] suggests that much larger scientific and technological networks (USPTO patents or the entire arXiv.org) can undergo similar topological transitions, even if they are not a single field of research. Thus the presence of a topological transition in a collaboration graph is likely only a necessary condition, but not a sufficient one, in terms of identifying scientific or technological breakthroughs. To establish (or falsify) this case it is therefore necessary that more evidence for more fields be collected and analyzed, and that the properties of the corresponding graphs' critical behavior be shown indeed to be general and characterized by the same exponents. It is also possible, as suggested by Kleinberg [14] and others [39,40,44-46], that simple patterns in text – for example, the anomalously high occurrence of certain words or word combinations – can be correlated with changes in network properties, helping to make the case for the simultaneous social and conceptual cohesiveness of a field promoted by the discovery of new concepts or techniques. Such a finding would help establish the generality of the cognitive and social mechanisms that underlie all processes of discovery and innovation and, in the process, give us universal models that capture the essence of their statistics and dynamics.

References

[1] T. S. Kuhn, *The Structure of Scientific Revolutions*, 2nd. ed. (Chicago: Univ. of Chicago Press, 1970 [1962]).

[2] D. J. de Solla Price, *Little Science, Big Science* (Columbia Univ. Press, New York, 1963)

[3] A. N. Tabah, Literature dynamics: Studies of growth, diffusion, and epidemics, *Annual Review of Information Science and Technology (ASIS)* **34**: 249–286 (1999).

[4] D. J. de Solla Price, Networks of scientific papers, *Science* **149**: 510—515 (1965).

[5] R. M. Shiffrin, K. Boerner, Mapping knowledge domains, *Proc. Natl. Acad. Sci. USA*, **98**: 5183–5185 (2001).

[6] K. W. Boyack, R. Klavans, K. Boerner, Mapping the backbone of science, *Scientometrics*, **64**: 351–374 (2005).

[7] J. Bollen, H. Van de Sompel, A. Hagberg, L. M. A. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva, Clickstream data yields high-resolution maps of science, to appear in *PLoS One*.

[8] W. Goffman, V. A. Newill, Generalization of epidemic theory: An application to the transmission of ideas, *Nature* **204**: 225–228 (1964).

[9] W. Gofman, Mathematical approach to the spread of scientific ideas: The history of mast cell research, *Nature* **212**: 449–452 (1966).

[10] W. Goffman, G. Harmon, Mathematical approach to the prediction of scientific discovery, *Nature* **229**: 103–104 (1971).

[11] E. Garfield, The epidemiology of knowledge and the spread of scientific information, *Current Contents* **35**: 5–10 (1980).

[12] L. M. A. Bettencourt, A. Cintron-Arias, D. I. Kaiser, C. Castillo-Chavez, The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models, *Physica A* **364**: 513–536 (2006).

[13] L. M. A. Bettencourt, D. I. Kaiser, J. Kaur, C. Castillo-Chávez, and D. E. Wojick, Population modeling of the emergence and development of scientific fields, *Scientometrics* **75**: 495-518 (2008).

[14] J. Kleinberg. Bursty and Hierarchical Structure in Streams. Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (2002).

[15] C. Chen, Searching for intellectual turning points: Progressive knowledge domain visualization, *Proc. Natl. Acad. Sci. USA* (suppl.) **101:** 5303–5310 (2004).

[16] A.F.J. Van Raan, On growth, ageing, and fractal differentiation of science, *Scientometrics* **47**: 347-362 (2000).

[17] M. E. J. Newman, Scientific collaboration networks: I. Network construction and fundamental results, *Phys. Rev. E* **64**: 016131 (2001).

[18] M. E. J. Newman, Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, *Phys. Rev. E* **64**: 016132 (2001).

[19] M. E. J. Newman, The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA* **101**: 404–409 (2004).

[20] S. Redner, Citations statistics from 110 years of Physical Review, *Physics Today* **58**: 49 (2005).

[21] H. Collins, *Changing Order: Replication and Induction in Scientific Practice*, 2nd ed. (Chicago: Univ. of Chicago Press, 1992 [1985]).

[22] B. Latour, *Science in Action* (Cambridge: Harvard Univ. Press, 1987).

[23] R. E. Kohler, *Lords of the Fly: Drosophila Genetics and the Experimental Life* (Chicago: Univ. of Chicago Press, 1994).

[24] P. L. Galison, *Image and Logic: A Material Culture of Microphysics* (Chicago: Univ. Chicago Press, 1997).

[25] D. Kaiser, *Drawing Theories Apart: The Dispersion of Feynman Diagrams in Postwar Physics* (Chicago: Univ. of Chicago Press, 2005).

[26] I. Lakatos and A. Musgrage, eds., *Criticism and the Growth of Knowledge* (New York: Cambridge Univ. Press, 1970).

[27] J. Rouse, *Knowledge and Power: Toward a Political Philosophy of Science* (Ithaca: Cornell Univ. Press, 1987).

[28] P. Horwich, ed., *World Changes: Thomas Kuhn and the Nature of Science* (Cambridge: MIT Press, 1993).

[29] M. Gerstein, and S M Douglas "RNAi Development", *PLoS Comput Biol* **3**: e80 (2007).

[30] B. Uzzi, and J. Spiro. "Collaboration and Creativity: The Small World Problem," *Am J. Soc* **111**: 447-504 (2005).

[31] J. Leskovec, J. Kleinberg, C. Faloutsos "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations" ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (2005).

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

[33] M. E. J. Newman. Who Is the Best Connected Scientist? A Study of Scientific Coauthorship Networks. *Lecture Notes in Physics* **650**: 337-370 (2004).

[34] J.J. Ramasco, S.N. Dorogovtsev, R Pastor-Satorras, Self-organization of collaboration networks, *Phys. Rev. E* **70:** 036106 (2004)

[35] K. Börner, L. Dall'Asta,, W. Ke and A. Vespignani, Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams, *Complexity* **10**: 57-67 (2005).

[36] R. Cohen, and S. Havlin, Scale-Free Networks Are Ultrasmall, *Phys. Rev. Lett.* **90**:058701 (2003)

[37] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Statistical mechanics of topological phase transitions in networks, *Phys. Rev. E* **69**: 046117 (2004).

[38] J. Zinn-Justin, *Quantum Field Theory and Critical Phenomena* (Oxford, UK:

University Press, 1989).

[39] D. Chavalarias, and J. Cointet, Science mapping with asymmetric co-occurrence analysis: Methodology and case study, In: *Proceedings of the European Conference on Complex Systems*, Dresden, 1–5 (2007).

[40] D. Chavalarias, and J. Cointet, Bottom-up scientific field detection for dynamical and hierarchical science mapping, methodology and case study, *Scientometrics* **75**: 37–50 (2008).

[41] T. Kuhn, Energy conservation as an example of simultaneous discovery, *Critical Problems in the History of Science*, ed. M. Clagett (Madison: Univ. of Wisconsin Press, 1959), 321-356.

[42] T. Arabatzis, *Representing Electrons: A Biographical Approach to Theoretical Entities* (Chicago: Univ. of Chicago Press, 2006).

[43] Committee on Science, Engineering, and Public Policy, *Enhancing the Postdoctoral Experience for Scientists and Engineers* (Washington, D.C.: National Academy of Sciences, 2000).

[44] R.R. Braam, H.F. Moed, and A. F. J. Van Raan, Mapping of science by combined cocitation and word analysis. II. dynamical aspects, *Journal of American Society for Information Science* **42**: 252–266 (1991).

[45] R. Buter, and E. Noyons, Using bibliometric maps to visualise term distribution in scientific papers, In: Sixth International Conference on Information Visualisation (IV'02), pp. 697–702 (2002).

[46] M. Callon, J. Courtial, and F. Laville, Co-word analysis as a tool for describing the network of interaction between basic and technological research: The case of polymer chemistry, *Scientometrics* **22**: 155–205 (1991).