

Collaboration, Stars, and the Changing Organization of Science

Evidence from Evolutionary Biology

Ajay Agrawal, John McHale, and Alexander Oettl

3.1 Introduction

The spatial organization of science is undergoing a fundamental transformation. New patterns of institutional participation, division of labor, and star scientist centrality are emerging. Given the essentially combinatorial nature of invention and innovation, changes in organization that affect access to knowledge and ease of collaboration to produce new knowledge are potentially of great importance to aggregate technological progress and economic growth.¹

In this chapter, we document and discuss significant changes in the spatial organization of science over recent decades in the field of evolutionary biology. Specifically, we identify two trends that appear contradictory at first glance. First, we find that the concentration of scientific output at the institution level is falling. More institutions are participating in scientific research over time and relatively more activity is migrating to lower-ranked institu-

Ajay Agrawal is the Peter Munk Professor of Entrepreneurship and associate professor of strategic management at the Rotman School of Management of the University of Toronto and a research associate of the National Bureau of Economic Research. John McHale is Established Professor and Head of Economics at the National University of Ireland, Galway. Alexander Oettl is assistant professor of strategic management at the Georgia Institute of Technology.

This research was funded by the Centre for Innovation and Entrepreneurship at the Rotman School of Management, University of Toronto and the Social Sciences and Humanities Research Council of Canada. We thank Adam Jaffe, Ben Jones, Julia Lane, and seminar participants at the NBER preconference workshop on “The Changing Frontier: Rethinking Science and Innovation Policy” as well as at the University of Toronto for valuable feedback. Errors remain our own. For acknowledgments, sources of research support, and disclosure of the authors’ material financial relationships, if any, please see <http://www.nber.org/chapters/c13038.ack>.

1. For influential work that emphasizes the role of combining ideas in the generation of new knowledge, see Romer (1990), Jones (1995), Weitzman (1998), and Mokyr (2002).

tions, broadening the base of science. At the same time, however, we find that the concentration of scientific output at the individual level is increasing. Publications and citations have always been highly skewed toward star performers. However, the relative importance of stars has increased in recent decades.

What could explain these seemingly contradictory trends? Collaboration offers a possible explanation. An increase in collaborative activity could broaden the base of science at the department level by raising the relative amount of participation by previously lesser-involved research institutions and at the same time increase the concentration of output at the individual level by disproportionately benefiting highly productive scientists, perhaps through more efficient matching with collaborators, thus enabling more finely grained specialization. Stars may disproportionately benefit from better matching because they have a larger pool of potential distant collaborators to choose from. We report descriptive evidence that is consistent with these conjectures.

Specifically, we report evidence that the level of collaboration has increased significantly over time. Furthermore, the average distance between collaborators has grown in terms of both physical distance and the rank separation of collaborating institutions, consistent with the conjecture that collaboration plays a role in the expanding base of science. We further show that the base of institutional participation has grown, including the entry of institutions from emerging economies. Moreover, we show that star scientists have an increasingly larger pool of potential collaborators relative to nonstars, consistent with the assertion that they may disproportionately benefit from lower communication costs due to better matching opportunities.

Why might collaboration be increasing? We see two central forces that increase the returns to collaboration although also work in opposing directions on the returns to colocation. The first is the rising “burden of knowledge” (Jones 2009). The increasing depth of knowledge required to work at the scientific frontier is leading to increasing returns to specialization. This in turn raises the returns to collaboration, given the need to combine ideas and skills to produce new ideas. Furthermore, to the extent that colocation lowers the cost of collaboration, the rising burden of knowledge increases the returns to colocation. Agrawal, Goldfarb, and Teodoridis (2013) report evidence consistent with the knowledge burden hypothesis. Utilizing the sudden and unexpected release of previously hidden knowledge caused by the collapse of the Soviet Union as a natural experiment, the authors find that an outward shift in the knowledge frontier does indeed cause an increase in collaboration.

The second force is the improvement in collaboration-supporting technologies that reduce the barriers created by distance such as e-mail, low-cost conferencing, and file-sharing technologies (Agrawal and Goldfarb 2008; Kim, Morse, and Zingales 2009). All else equal, these advances allow

for a greater physical dispersal of collaborating scientists. So, although the declining cost of communication may decrease the returns to colocation, it increases the returns to collaboration. Therefore, despite the potential conflict between these two forces with respect to their impact on the relative returns to colocation, they both increase the returns to collaboration.

We develop a simple model that provides a potential unified explanation for these facts. The key idea behind the model is that stars have a larger set of potential collaborators to choose from—perhaps due to having more former graduate students—and thus have the potential to gain disproportionately from improvements in collaboration technology. Moreover, some of these cross-institutional collaborations may occur with scientists from lesser-ranked institutions, consistent with a broadening institutional base in the production of science.

Drawing on parallel work on the causal impact of star scientists on departmental performance (Agrawal, McHale, and Oettl 2013), we speculate on the efficiency of the emerging spatial distribution of scientific activity. Recognizing the existence of knowledge, reputational, and consumption externalities associated with the location decisions of star scientists, we make no presumption that the resulting spatial distribution of stars is efficient. We find that stars attract other stars and also that the recruitment of a star can have positive effects on the productivity of colocated incumbents working in areas related to the star. These effects appear to be particularly strong when recruitment takes place at non-top-ranked institutions.

However, while strong forces may lead to star agglomeration due to localized externalities, lower-ranked institutions may have strong incentives to compete for stars as a core part of strategies aimed at climbing the institutional rankings. We document significant movement up the rankings for a select set of institutions that begin outside the top-ranked institutions. Congestion effects may also exist from star colocation due to clashing egos and increasing returns to “vertical collaboration” across skill sets located at different institutions. Overall, we find a tendency toward reduced concentration of the field’s best scientists at its top-ranked institutions. In addition, the increasing propensity to collaborate across institutions, particularly across institutions of significantly different rank, further diminishes the concentration of knowledge production. Thus, fears of excessive concentration of stars due to positive sorting might be overblown, although research on the normative implications of the observed changes in the organization of science is still at an early stage.

We structure the rest of the chapter as follows. In section 3.2, we explain the construction of our evolutionary biology data at the institutional and individual levels. In section 3.3, we report evidence of the broadening institutional and international base of scientific activity. We describe the increasing concentration of individual productivity and, in particular, the rising importance of stars in section 3.4. In section 3.5, we present data on the

overall rise of collaborative activity and the change in collaboration patterns. In section 3.6, we develop a model that offers a potential unified explanation of the facts documented in previous sections. Finally, in section 3.7 we provide a more speculative discussion of possible normative implications of these participation, concentration, and collaboration patterns, with an emphasis on the role of the location of stars and the increasing propensity to collaborate across institutions.

3.2 Data

Our study focuses on the field of evolutionary biology, a subfield of biology concerned with the processes that generate diversity of life on earth. Although some debate exists among historians of science and practicing evolutionary biologists over the key early contributors to this discipline, the general consensus remains that *On the Origin of Species by Means of Natural Selection*, authored by Charles Darwin in 1859, is the foundational text of this field. As in most fields of science, research in evolutionary biology consists of both theoretical and experimental contributions. In addition to specializing in particular topic areas, empiricists often specialize in working with particular organisms such as *Macrotrachela quadricornifera* (rotifer), *Drosophila melanogaster* (fruit fly), and *Gasterosteus aculeatus* (three-spined stickleback fish). The returns to species specialization result from, for example, the upfront costs of learning how to work with a particular species (including, in many cases, learning where to find them and how to catch and care for them to facilitate reproduction in order to observe, for instance, the variation in genotypes and phenotypes of offspring over multiple generations) as well as setting up the infrastructure in a lab or in nature to study them.

3.2.1 Defining Evolutionary Biology

Defining knowledge in evolutionary biology is not straightforward. On the input side, evolutionary biology, as in many areas of science, draws from many fields, such as statistics, molecular biology, chemistry, genetics, and population ecology. Furthermore, on the output side, some of the most influential papers are published in general interest as opposed to field-specific journals. Therefore, identifying the set of papers that comprise the corpus of the field is complicated because although every paper in the *Journal of Evolutionary Biology* is probably relevant, most papers in *Science* and *Nature* are not, although a significant fraction of the field's most important papers are published in those latter two journals.

Therefore, we follow a three-step process for defining “evolutionary biology papers.” First, using bibliometric data from the ISI Web of Science, we collect data on all articles published during the twenty-nine-year period of 1980 through 2008 in the journals associated with the four main societies that focus on the study of evolutionary biology: the Society for the Study

of Evolution, the Society for Systematic Biology, the Society for Molecular Biology and Evolution, and the European Society of Evolutionary Biology. Their respective journals are: *Evolution*, *Systematic Biology*, *Molecular Biology and Evolution*, and *Journal of Evolutionary Biology*. We focus on these four society journals because every article published within them is relevant to evolutionary biologists. In other words, unlike general interest journals such as *Science*, *Nature*, and *Cell*, which include papers from evolutionary biology but also research from many other fields, these four journals focus specifically on our field of interest. This process yields 15,256 articles.

Second, we collect all articles that are referenced at least once by these 15,256 society journal articles. There are 149,497 unique articles that are referenced at least once by the set of 15,256 evolutionary biology society articles. This set of 149,497 articles includes, for example, papers that are important to the field but are published outside the four society journals, such as key evolutionary biology papers published in *Science* that are cited, likely multiple times, by articles in the four society journals. We call this set of 149,497 papers the corpus of influence because each of these articles has had impact on at least one “pure” evolutionary biology article.

Third, we citation-weight the corpus of influence. We do this by counting the references to each of the 149,497 articles from the original 15,256 society journal articles. There are 501,952 references from the 15,256 society journal articles. So, on average, articles in the corpus are cited 3.4 times. Unsurprisingly, the distribution of citations is highly skewed. The minimum number of citations is one (by construction), the median is one, and the maximum is 906.² For most of the analyses in this chapter, we use counts of citation-weighted publications. When we do so, we use the 149,497 articles weighted by the 501,952 society article references.

3.2.2 Identifying Authors

We follow several steps to attribute the 149,497 articles in the corpus of influence to individual authors. The reason this process requires several steps is that authors are not uniquely identified and therefore name disambiguation is necessary. In other words, when we encounter multiple papers authored by James Smith, we need to determine whether each is written by the same James Smith or if instead these are different people with the same name. This process is made more challenging because until recently the ISI Web of Science only listed the first initial, a middle initial (if present), and the last name for each author. Is J Smith the same person as JA Smith? Name disambiguation is particularly important for properly assessing researcher productivity over time and changing collaboration patterns.

2. This paper is “The Neighbor-Joining Method—A New Method for Reconstructing Phylogenetic Trees” published in *Molecular Biology and Evolution* (1987) by Saitou Naruya (University of Tokyo) and Masatoshi Nei (University of Texas).

To address this issue, we employ heuristics developed by Tang and Walsh (2010). The heuristic utilizes backward citations of focal papers to estimate the likelihood of the named author being a particular person. For example, if two papers reference a higher number of the same papers (weighted by how many times the paper has been cited, i.e., how popular or obscure it is), then the likelihood of those two papers belonging to the same author is higher. We attribute two papers to the same author if both papers cite two or more rare papers (fewer than fifty citations) in both papers. We repeat this process for all papers that list nonunique author names (i.e., same first initial and last name). We exclude scientists who do not have more than two publications linked to their name.

Overall, 171,428 authors are listed on the 149,497 articles. We drop 140,240 names because they do not have more than two publications linked to their name. Employing the process described above, we assign the remaining 31,188 author names to 32,955 unique authors (a single name may map to more than one person). We conduct our analyses using these 32,955 authors. It is important to note that this is the total number of scientists in our sample over the twenty-nine-year period, but that the number of active scientists varies from year to year. Unsurprisingly, the output produced by these authors is highly skewed. Considering the overall period of our study, the minimum number of publications per author is three (by construction), the median is four, the mean is 7.5, and the maximum is 210 (Professor Rick Shine at the University of Sydney).

We use citation-weighted paper counts per year as our primary measure of author output. We treat as equal every paper on which a scientist is listed as an author. In other words, we do not distinguish between a paper on which a scientist is one of two authors from one on which they are one of three authors. An alternative approach is to use fractional paper counts where half a paper unit is attributed to the focal author in the former case and a third in the latter. Although we report results using the former approach, we conduct our analysis using both approaches. The results are qualitatively similar.

In certain analyses, we refer to the top 100 (or 200, or 50) scientists. When we do so, we determine ranking by the accumulated stock of citation-weighted output over the preceding years. When we refer to “stars,” we are referring to scientists in the 90th percentile in a given year in terms of their accumulated stock of citation-weighted paper output over the preceding years. We provide a more detailed explanation of how we identify stars and related features of the data in our companion paper that focuses on stars and that uses the same data (Agrawal, McHale, and Oettl 2013).

3.2.3 Identifying Scientist Locations

Using the unique author identifiers we generate in the process described above for each evolutionary biology paper, we then attribute each scientist to a particular institution for every year they are active. A scientist is active

from the year they publish their first paper to the year they publish their last paper. Here again, we must overcome a data deficiency inherent within the ISI Web of Science data; until recently, the Web of Science did not link institutions listed on an article to the authors. Instead, we impute author location using reprint information that provides a one-to-one mapping between the reprint author and the scientist's affiliation. In addition, we take advantage of the fact that almost 57 percent of evolutionary biology papers are produced with only a single institution listing. Thus, we are able to directly attribute the location of all authors on these papers to the focal institution. This method of location attribution is more effective for evolutionary biology than for many other science disciplines since articles in this field are generally produced by smaller-sized teams relative to other disciplines in the natural sciences (3.32 average number of authors per paper).

Overall, we are able to attribute 78.9 percent of the 32,955 unique authors to an institution. We drop institutions that do not produce at least one publication in each of the twenty-nine years under study. This results in the identification of 255 institutions that actively produce new knowledge in the field of evolutionary biology throughout our study period. Although we refer to these as "departments," they are actually the set of authors at an institution (e.g., Georgia Tech) who publish at least three articles that we categorize as being part of the corpus of influence in evolutionary biology during the study period. In other words, these individuals may not all formally belong to the same department within the institution. Again, the output of departments is highly skewed. Over the twenty-nine-year period, the minimum number of publications per institution per year is one (by construction), the median is eleven, the mean is 17.7, and the maximum is 181 (Harvard University in 2005).

3.3 Participation: A Broadening Base

The first trend in the organization of evolutionary biology we document is a decline in the skew of the distribution of output across institutions. This may reflect: (a) an increasing emphasis in knowledge production across previously lesser-producing institutions that are now more concerned about rankings and thus increasingly emphasizing research output as a factor in promotion and tenure, (b) changing preferences of faculty who have spent more time than their predecessors developing specialized research expertise, and/or (c) mounting political pressure to distribute government funding more evenly across institutions and political jurisdictions. In addition, we find a dramatic increase in research activity in emerging economies, possibly reflecting a broader movement toward higher value-added activities as part of the economic development process.

In figures 3.1, 3.2, and 3.3, we report evidence of the broadening base of science in terms of the declining department-level concentration of sci-

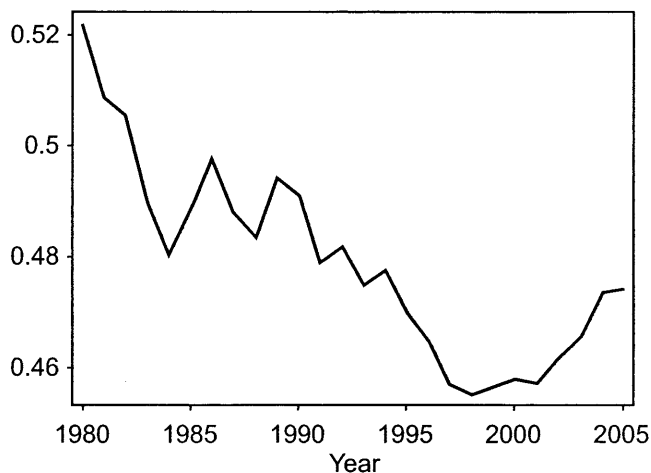


Fig. 3.1 Gini coefficients by year for the distribution of scientists across departments

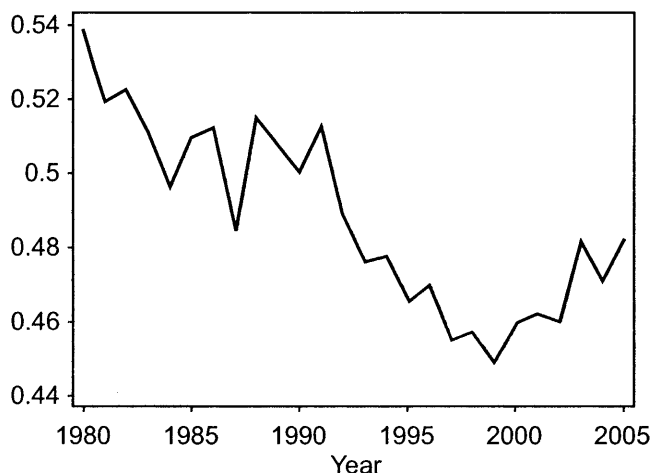


Fig. 3.2 Gini coefficients by year for the distribution of publications across departments

entists, publications, and citations, respectively. Specifically, we plot Gini coefficients to illustrate the distribution of scientists (publications, citations) across departments by year. The pattern of falling concentration is pronounced for the period between 1980 and 2000, although there is some indication of a turnaround in this pattern after 2000.

In figure 3.4, we plot department-level Lorenz curves for publications and citations. These curves illustrate the overall shift in the distribution over time. For example, the top 20 percent of departments produce 60 percent of all publications in 1980, but only 50 percent in 2000. Similarly, the

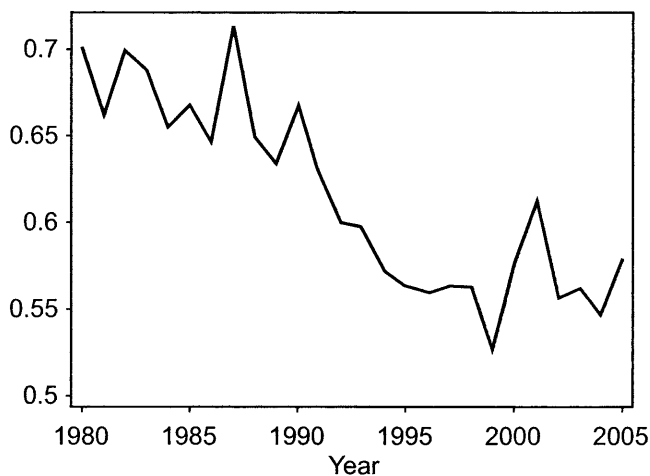


Fig. 3.3 Gini coefficients by year for the distribution of citations received across departments

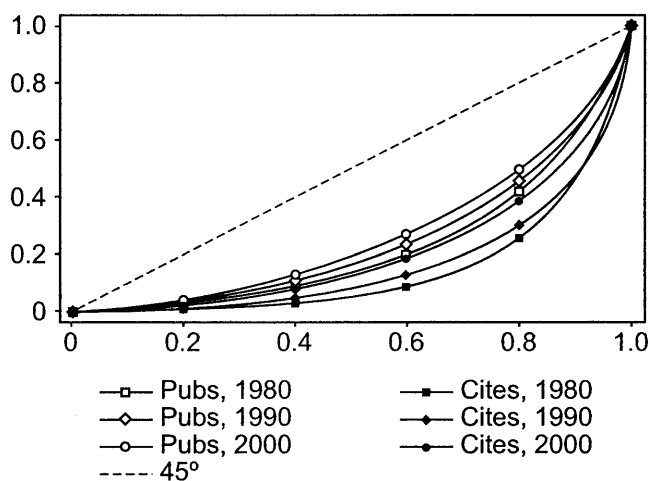


Fig. 3.4 Lorenz curves by department for publications and citation-weighted publications

top 20 percent produce 75 percent of all citation-weighted publications in 1980, but only 60 percent in 2000. It is important to note that we use a balanced panel for these analyses, including only the 255 institutions publishing in evolutionary biology throughout the period under study. In other words, we do not allow for entry of new institutions part way through the study period. Since most institutions that are ever meaningful contributors to this field are active throughout our study period, this is not a serious restriction.

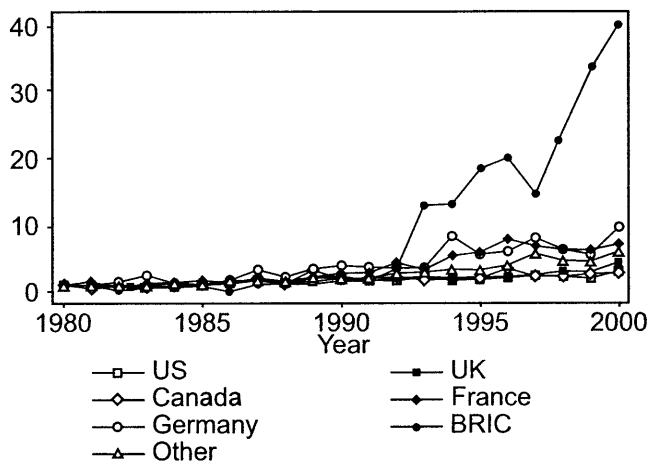


Fig. 3.5 Publication count by country by year normalized relative to output in 1980

However, we relax the no-entry restriction for the data we use in the next graph where we plot output by country because several institutions in previously low-income countries are not active in the early years, but have since become increasingly important in the overall production of knowledge. We plot the increasing importance of institutions based in emerging markets in figure 3.5. The growth rate of publications from institutions based in BRIC countries (Brazil, Russia, India, and China) begins to rise dramatically from the early 1990s onward and increases fortyfold by 2000. However, in absolute terms, the BRIC countries are still minor knowledge producers in this field relative to the leading nations, such as the United States, the United Kingdom, France, Germany, and Canada.

These decentralization findings from university-based research in evolutionary biology are consistent with prior findings on the decentralization of innovative activity more broadly (Rosenbloom and Spencer 1996; Bresnahan and Greenstein 1999). Also, more recently, in a study of innovation in information and communication technologies (ICTs) over almost the identical period as our study (1976–2010), Ozcan and Greenstein (2013) examine US patent data and find that although the top twenty-five firms account for 72 percent of the entire patent stock and 59 percent of new patents in 1976, they account for only 55 percent and 50 percent, respectively, by 2010. The decline is even more dramatic when they restrict the sample to the ownership of high-quality patents (82 percent down to 62 percent). They interpret their results as supporting the view that decentralization is resulting from “more widespread access to the fundamental knowledge and building blocks for innovative activity” (5).

Overall, we interpret our data as reflecting a decline in the concentration of output at the department level. In other words, the top institutions

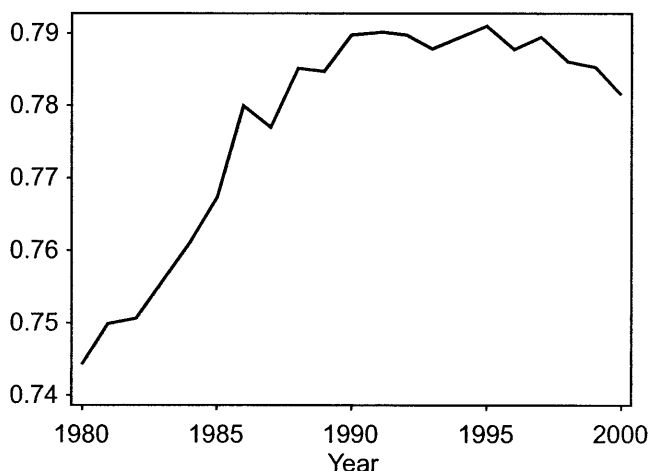


Fig. 3.6 Gini coefficients for the distribution of citation-weighted publications across individuals

are producing a decreasing fraction of the overall output, and previously lesser-producing institutions are now contributing a higher portion of overall output. However, this is not the case at the individual level. We turn to this unit of analysis next.

3.4 Concentration: The Increasing Importance of Stars

With greater democratization in knowledge production across departments, is science becoming a less elite activity, with a falling centrality of stars as they compete with scientists from an ever-widening base? One might expect the broadening base of science at the department level to be accompanied by a reduction in the concentration of output at the individual level. However, we find evidence of the opposite.

We again plot Gini coefficients by year using citation-weighted publications, but this time at the individual level. These data, illustrated in figure 3.6, indicate a significant increase in concentration during the 1980s and then relative stability during the following decade. Then, in figure 3.7, we plot individual-level Lorenz curves for 1980, 1990, and 2000 with the same data to illustrate how the full distribution shifts over time. Again, we see individual-level output increasing over time. For example, the top 20 percent of scientists produce 70 percent of output in 1980 but 80 percent by 2000, with most of the shift occurring in the first decade. Furthermore, in figure 3.8, we illustrate the increasing spread between the top-performing scientists and the rest by comparing the number of citation-weighted publications required to be in the top 50, which increases fivefold, to the average number

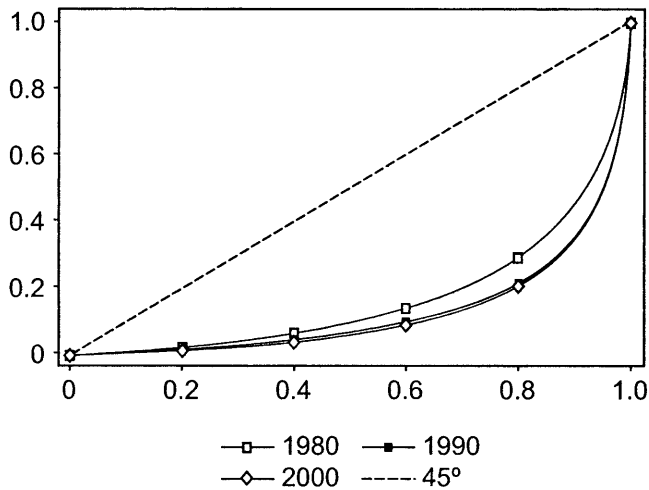


Fig. 3.7 Lorenz curves by individual for citation-weighted publications

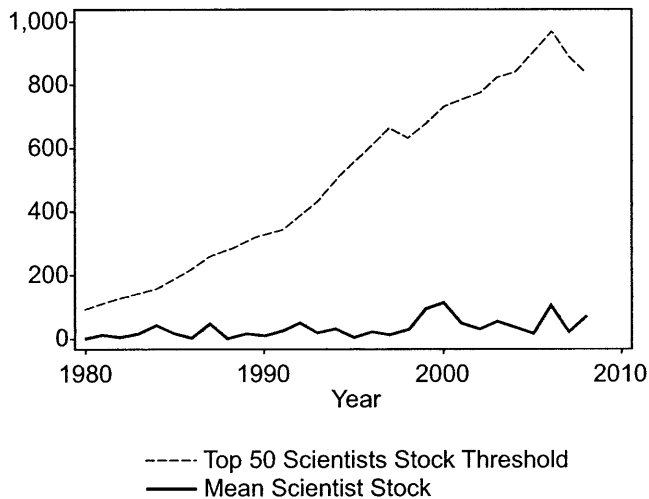


Fig. 3.8 Publication stock of 50th-ranked scientist

of citation-weighted publications, where the increase over the same time period is negligible.

How might we reconcile decreasing concentration at the department level but increasing concentration at the individual level? The answer may lie in the changing patterns of collaboration. Recall that although the rising burden of knowledge and declining communication costs exert opposing forces on the returns to colocation, both increase the returns to collaboration. We turn to the topic of collaboration next.

3.5 Collaboration: Increasing across Distance and Rank

The trend toward increasing collaboration is a well-documented feature of the changing organization of science (e.g., see Wuchty, Jones, and Uzzi 2007). In fact, the rising role of collaboration is one of the most common themes across the chapters in this volume. For example, Branstetter, Li, and Veloso (chapter 5) state: “Our study suggests that the increase in US patents in China and India are to a great extent driven by MNCs (multinational corporations) from advanced economies and are *highly dependent on collaborations* with inventors in those advanced economies.”³ Forman, Goldfarb, and Greenstein (chapter 6) explain: “We show that these [geographic distribution of inventive activity] results are *largely driven by patents filed by distant collaborators* rather than by noncollaborative patents or by patents by nondistant collaborators.” Stephan (chapter 10) argues: “Much of the equipment associated with these shifts in logic were, although expensive, still affordable at the lab or institutional level. Some, however, such as an NMR (nuclear magnetic resonance), carried sufficiently large price tags to *encourage, if not demand, collaboration across institutions*.” Conti and Liu (chapter 2) report: “Collaborations with other scientists, as measured by the number of coauthors on a paper, have increased. This increase is driven by *collaborations with scientists outside of a trainee’s laboratory*.” Freeman, Ganguli, and Murciano-Goroff (chapter 1) discover through their survey: “*The major factor cited for all types of collaborations was ‘unique knowledge, expertise, capabilities. . . . Non-colocated and international teams were more likely to have a coauthor contributing data, material, or components—a pattern that has been increasing over time.*”

We document this phenomenon of increasing collaboration over time in our own setting in figure 3.9. Specifically, this figure illustrates the steady increase in the average number of authors on evolutionary biology papers, rising from 2.3 in 1980 to 3.8 in 2005. Moreover, this collaboration increasingly has been taking place across university boundaries (Jones, Wuchty, and Uzzi 2008). We illustrate this in figure 3.10, where we plot the average number of unique institutions represented on a paper over time. The figure shows that this number increases from 1.46 in 1980 to 2.45 in 2005.

We also observe a dramatic trend in the average rank difference between authors on coauthored papers (figure 3.11). For example, in 1980 the average distance in rank between collaborating institutions is approximately thirty (e.g., one collaborator is at an institution ranked number twenty and the coauthor is at an institution ranked number fifty), by 2005 the difference increases to approximately fifty-five. Furthermore, we find evidence of increasing distance between collaborators over time. We illustrate this in

3. The emphasis in this and the other quotes in this paragraph is our own, not that of the original authors.

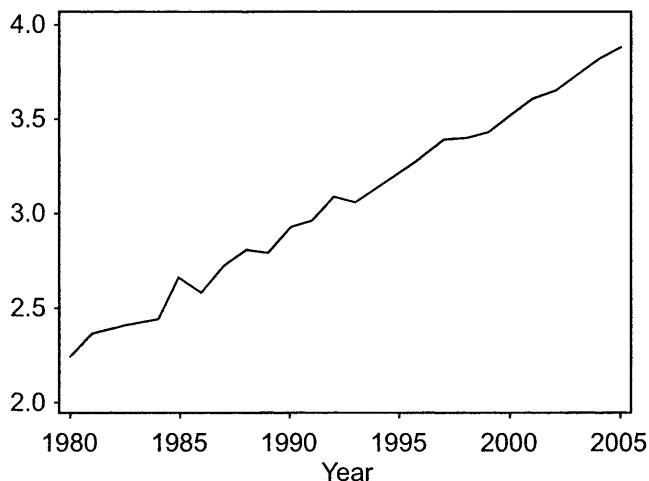


Fig. 3.9 Mean number of authors per paper

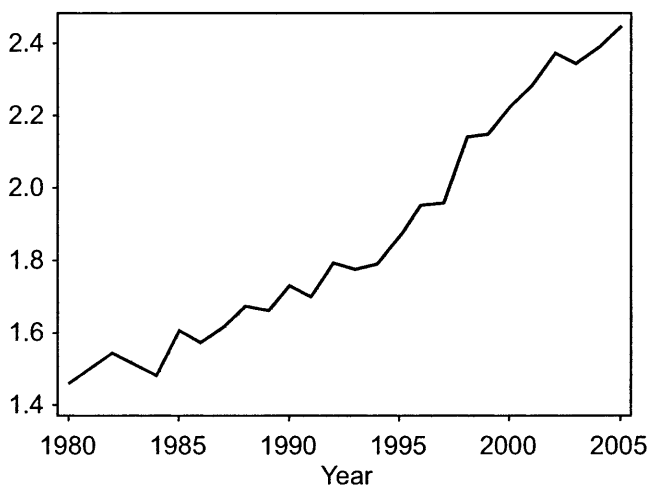


Fig. 3.10 Mean number of unique institutions per paper

figure 3.12, where the average distance between coauthors increases from 325 to 500 miles over the period 1980 to 2005.

Why might the falling cost of distant collaboration disproportionately benefit stars? Freeman, Ganguli, and Murciano-Goroff (chapter 1, this volume) present survey evidence indicating that, in general, a large fraction of collaborations occur between scientists who were previously colocated. We conjecture that one reason stars disproportionately benefit from a drop in the cost of distant collaboration is because they have a greater number of distant

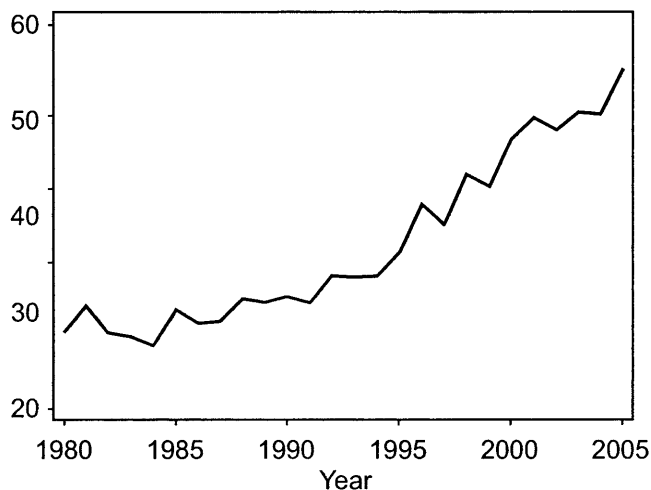


Fig. 3.11 Mean difference in institution rank between coauthors

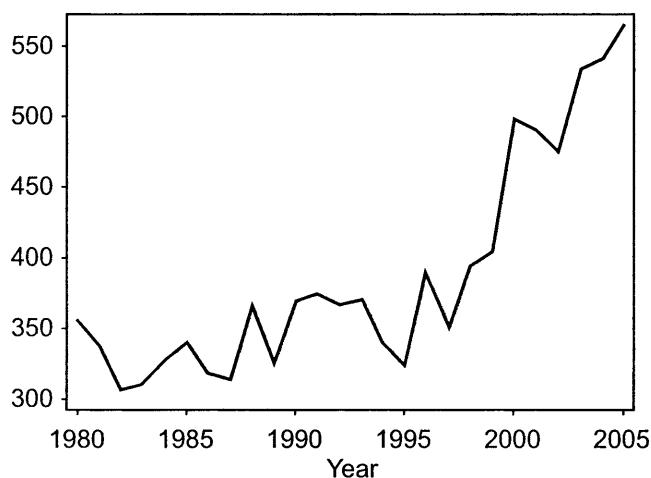


Fig. 3.12 Mean distance between coauthors (miles)

potential collaborators. For example, stars are likely to have more graduate students and postdoctoral students than nonstars, on average, and these students are likely to subsequently move to other institutions. To the extent that communication technologies like the Internet are most suitable for facilitating communication between individuals with an already established relationship as opposed to establishing new relationships (Gaspar and Glaeser 1998), then lowering communication costs will disproportionately benefit those individuals, such as stars, who have more previously colocated, but now distant poten-

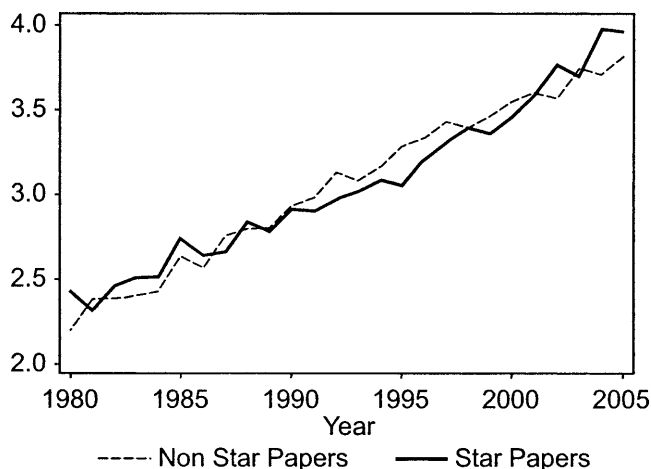


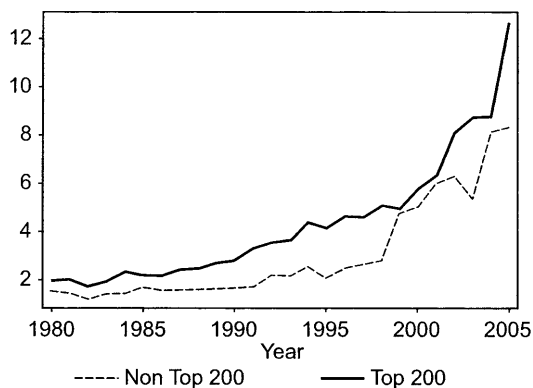
Fig. 3.13 Mean number of authors per paper (star versus nonstar)

tial coauthors. In other words, stars are able to employ this technology over a larger number of previously colocated but now distant potential collaborators.

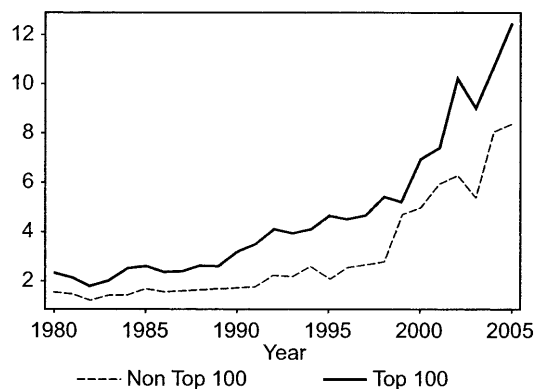
This benefit to stars could accrue through two non-mutually exclusive channels. First, stars could disproportionately increase the number of individuals they collaborate with. Our descriptive evidence suggests that although stars do increase their propensity to collaborate over time, so do nonstars. We illustrate this in figures 3.13 and 3.14. First, we show that although the number of coauthors per paper increases over time, there is no meaningful difference between papers with and without stars. Second, we construct three measures of stars (top 50, top 100, and top 200 scientists) and plot the number of unique coauthors per year for stars versus nonstars. These data indicate that although the annual number of unique collaborators is increasing over time for star scientists, stars do not seem to increase their number of unique collaborators at a meaningfully faster rate than nonstars.

Second, stars may disproportionately benefit from the fall in communication costs because they are able to make better matches with coauthors since they have more potential collaborators to choose from. In other words, the best of the available pool of potential collaborators is better for stars than for nonstars. So, for example, if the falling cost of communications increases the returns to collaboration such that both a star and a nonstar increase their number of collaborations by one, then the average star may choose the best-suited collaborator from a pool of many previously colocated but now distant potential collaborators, while the nonstar can only choose from a pool of few. Even if stars and nonstars are choosing collaborators from pools with the same distribution in terms of quality or range of skills, stars likely will be able to choose a superior match simply due to the larger pool size to which they have access.

(a) Top 200 Stars



(b) Top 100 Stars



(c) Top 50 Stars

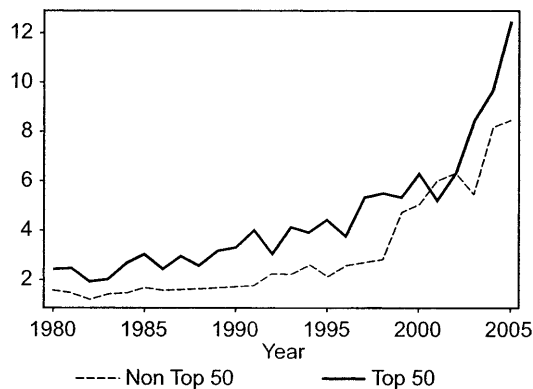


Fig. 3.14 Number of unique coauthors per year

Note: Panel (a) = top 200 stars; panel (b) = top 100 stars; and panel (c) = top 50 stars.

We construct a measure of the size of the pool of previously colocated but now distant collaborators by counting the cumulative number of individuals who coauthor with the focal scientist at least once while located at their home institution and then subsequently at least once while at another institution. We again construct three measures of stars (top 50, top 100, and top 200 scientists). In figure 3.15, we plot the potential distant coauthor pool size for stars versus nonstars (cumulative number of unique coauthors that were previously colocated but are now distant). It is important to note that this count is not simply the aggregation of the annual counts plotted in the prior figure. That is because in the prior figure repeated coauthorships are counted as distinct in each new year (although multiple coauthorships with the same individual in the same year are not double counted). However, in this plot only unique coauthorships that are unique in the absolute sense (cumulatively) are counted. Furthermore, in figure 3.15 we only count distant coauthors that were previously located whereas in the prior figure there were no distance or prior colocation restrictions in counting unique coauthors. These data indicate that the pool size of potential collaborators (such as graduate students and postdocs) grows significantly faster for stars than for nonstars. Furthermore, in figure 3.16, we plot the inverse of the ratio of the number of actual collaborations in a given year to the number of potential collaborators in the pool that year and compare the change in this ratio over time for stars versus nonstars. We interpret the ratio as a proxy for the degree of selectivity afforded to stars and nonstars. In other words, a higher ratio for stars versus nonstars indicates that stars collaborate with a smaller fraction of their pool of potential coauthors than nonstars. The figure thus suggests that the relative selectivity of stars versus nonstars in terms of choosing collaborators is increasing over time. While not conclusive, these descriptive data are consistent with the conjecture that stars disproportionately benefit from falling communication costs by way of an increased pool size of distant collaborations to choose from relative to nonstars.

3.6 Improved Collaboration Technology and the Distribution of Scientific Output: An Integrating Model

In this section, we develop a simple model to examine the effects of an improvement in collaboration technology on the distribution of scientific output. In particular, we examine how such an improvement both disproportionately affects stars and leads to more collaboration. The model's results are consistent with both an increased concentration of scientific output across individual scientists—that is, a star concentration effect—and also a broadening institutional base of science.

A key assumption is that relationships with previously colocated but now distant former coauthors, such as former graduate students and postdocs, are central to developing opportunities for subsequent collaboration. This is consistent with the survey evidence on collaboration reported by Freeman,

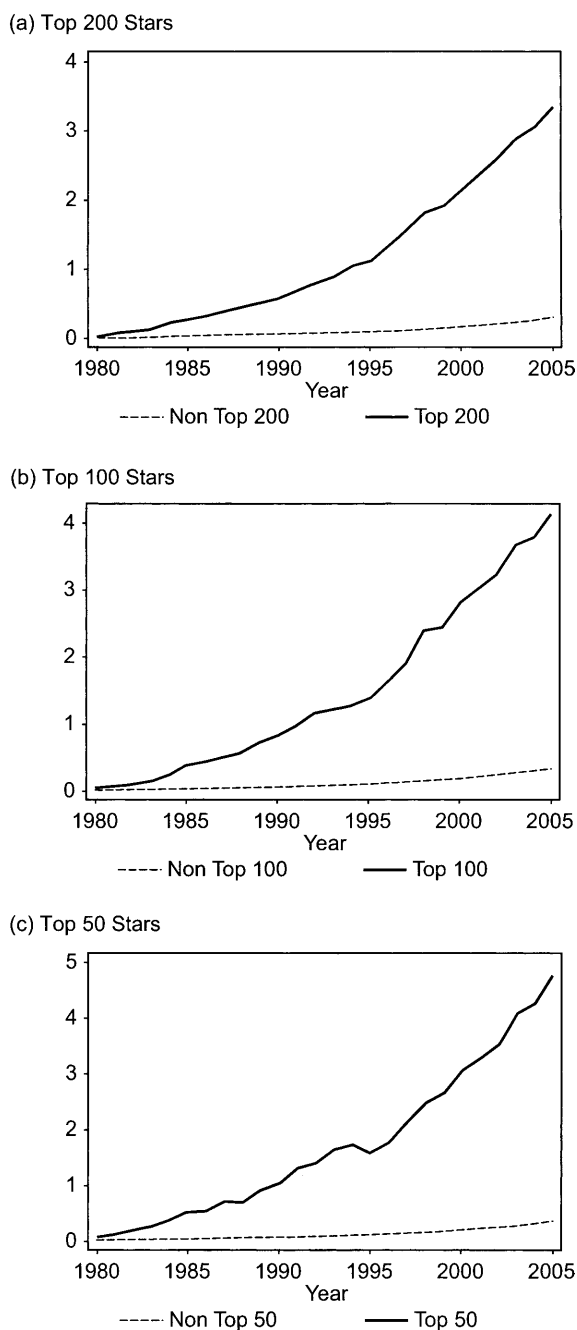


Fig. 3.15 Cumulative number of unique previously colocated but now distant coauthors

Note: Panel (a) = top 200 stars; panel (b) = top 100 stars; and panel (c) = top 50 stars.

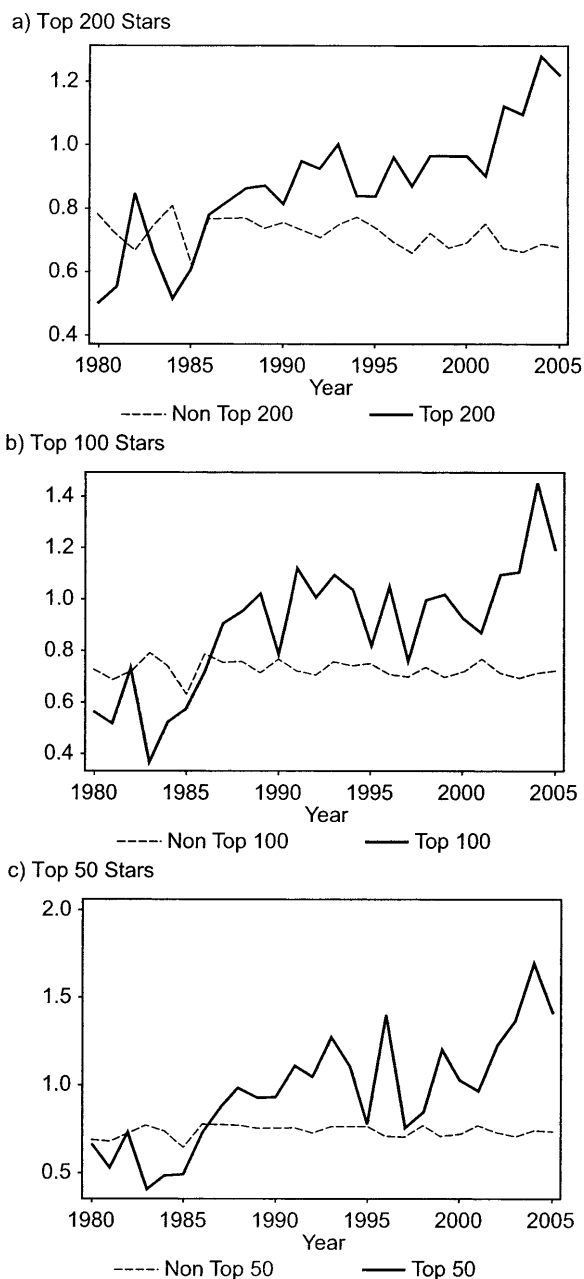


Fig. 3.16 Selectivity index

Note: Panel (a) = top 200 stars; panel (b) = top 100 stars; and panel (c) = top 50 stars. We construct the selectivity index as the inverse of the ratio of the number of unique collaborators in a given year over the cumulative number of previously colocated but now distant collaborators.

Ganguli, and Murciano-Goroff (chapter 1, this volume), which documents the extent to which such relationships account for the majority of collaborative partnerships. We also assume that the number of feasible collaborative relationships is limited due to the costs of collaboration. Furthermore, we assume that stars know a larger set of former graduate students and postdocs from which to choose their collaborative relationships. We do not need to assume that stars have better graduate students in general or engage in more collaborative relationships. We show that simply having a greater range of graduate students to choose from enables stars to gain disproportionately from an improvement in collaborative technologies, which we take to be due to improvements in communication technologies (e-mail, file-sharing technologies, etc.).

For a given scientist, we assume the value of a collaborative relationship, X , with a given former graduate student is uniformly distributed on the interval $[0, M]$. We assume that an improvement in collaboration technology increases the value of any relationship by a multiplicative factor. The increased value of collaboration could also reflect a greater need for collaboration due to the burden of knowledge effect. Thus, we can simply model an improvement in technology (or the greater need for collaboration) as an increase in M , effectively a stretching of the distribution to the right.

3.6.1 Basic Model

We assume initially that each scientist chooses the single-best relationship from her set of n former graduate students. We use the size of n as a proxy for the scientist's degree of stardom. For a given scientist, the expected value of the best available relationship is:

$$(1) \quad E(X) = \int_0^M X \frac{n}{X} \left(\frac{X}{M}\right)^n dX = \frac{n}{1+n} M.$$

This result uses the distribution of the maximum value of n , which draws from the uniform distribution.⁴

The increase in expected value from a small increase in the available collaboration technology is then:

$$(2) \quad \frac{\partial E(X)}{\partial M} = \frac{n}{1+n}.$$

The size of this increase is increasing in n ,

$$(3) \quad \frac{\partial^2 E(X)}{\partial M \partial n} = \frac{1}{(1+n)^2} > 0.$$

Thus, stars—those with a high n —gain disproportionately from the improvement in the collaboration technology.

4. The CDF for this extreme value distribution is: $F(X) = (X / M)^n$. The density function is then: $f(X) = (n / X)(X / M)^n$.

3.6.2 Extended Model

A limitation of the basic model is that it assumes a scientist will choose to collaborate with her best former graduate student no matter how low the value of that best collaboration. A more realistic assumption is that scientists have some threshold below which they will not collaborate, given the opportunity costs of collaboration (e.g., reduced time for sole authorship). Denoting this threshold as X^* , the expected value of a collaboration is now:

$$(4) \quad E(X) = \int_{X^*}^M \frac{n}{X} \left(\frac{X}{M} \right)^n dX = \left[\left(\frac{n}{1+n} \right) M \right] \left[1 - \left(\frac{X^*}{M} \right)^{n+1} \right].$$

The expected value is lower than when the threshold is absent because best draws from the distribution that are below the threshold result in zero value. It is also increasing in M , so that improvements in the collaboration technology are again beneficial.

$$(5) \quad \frac{\partial E(X)}{\partial M} = \left[\frac{n}{1+n} \right] \left[1 + n \left(\frac{X^*}{M} \right)^{n+1} \right] > 0.$$

We again ask if the technology improvement disproportionately benefits stars. This requires that the cross derivative with respect to n is positive. Making use of logarithmic differentiation, the cross derivative is:

$$(6) \quad \frac{\partial^2 E(X)}{\partial M \partial n} = \frac{1}{(1+n)^2} + \left(\frac{n^2}{1+n} \right) \left(\frac{X^*}{M} \right)^{n+1} \left[\frac{2}{n} - \frac{1}{1+n} + \ln \left(\frac{X^*}{M} \right) \right].$$

This cross derivative is obviously quite a complex function of n , M , and X^* . However, it can be shown to be positive for all n given a low enough value of X^* relative to the starting value of M , so that $\partial E(X) / \partial M$ is then monotonically increasing in n . Figure 3.17 shows the cross derivative as a function of n for different values of X^* (conveniently scaled by the starting value of M): 0.1, 0.2, and 0.3. At high values of X^* / M , the cross derivative can be negative over an intermediate range of n but becomes positive for high enough values of n . We assume, however, that the threshold is sufficiently low such that the cross derivative is positive for all n .

An additional consequence of introducing a threshold for collaboration is that the probability of collaboration is now itself a function of M .

$$(7) \quad \text{Prob}[X > X^*] = 1 - \left(\frac{X^*}{M} \right)^n.$$

This probability is also increasing in M , so that improvements in the collaborative technology lead to more as well as higher expected value collaboration:

$$(8) \quad \frac{\partial \text{Prob}[X > X^*]}{\partial M} = \frac{n}{M} \left(\frac{X^*}{M} \right)^n > 0.$$

Summing up, the extended model demonstrates two effects of an improvement in collaborative technology that could impact the distribution of scien-

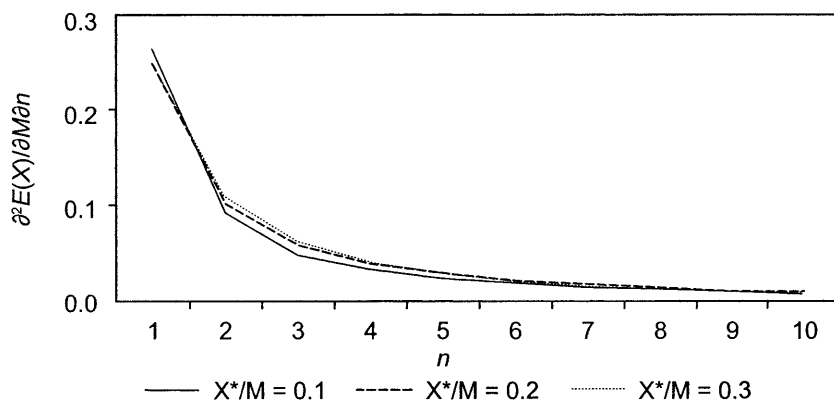


Fig. 3.17 Relationship of the cross derivative to n

tific output. First, provided that scientists do not set too high a threshold for engaging in collaboration, the benefit from the improvement in technology is increasing in n , so that stars—whom we assume to be disproportionately endowed with previously colocated but now distant former coauthors—benefit disproportionately. This is consistent with an increased concentration of scientific output at the individual level. Second, it will be beneficial for more scientists to engage in collaborative research. This is consistent with an expanding institutional base of science as more former students and postdocs—who will have dispersed across the institutional ranks—are involved in collaborative research.

3.7 Discussion: Normative Implications of Star Location

Our review of the basic trends in participation, concentration, and collaboration reveals the dramatically changing organization of scientific activity in the field of evolutionary biology. The emerging picture also points to the increasingly central role played by stars in collaboration and overall output. Moreover, stars, like the overall research community, appear to be increasingly collaborating across distance and institution rank. Overall, we see evidence of a developing cross-institutional division of scientific labor, with stars playing a leadership role in institution- and distance-spanning multiauthor research teams.

The rising centrality of stars raises questions about the efficient distribution of stars across institutions. We thus reflect on the efficiency of the emerging pattern of the division of labor, drawing on both the factual picture just documented and parallel work on the causal impact of star scientists at the departmental level (Waldinger 2012, 2013; Agrawal, McHale, and Oettl 2013). A key question is whether the emerging spatial distribution of stars is efficient from the perspective of maximizing the value of scientific output.

We do not presume that the distribution will be efficient, given the free location choices of individual scientists and the productivity, reputational,

and consumption externalities associated with those choices. We note in particular that the reputational spillover from locating at top-ranked institutions could lead to an excessive positive sorting of stars at these institutions. Such inefficiency, if it exists, could be ameliorated by easier cross-institution collaboration, effectively making the location of stars less important to knowledge production. Even so, given the ongoing costs of distance-related collaboration, a concern still remains that there may be excessive concentration from a social welfare perspective.

In Agrawal, McHale, and Oettl (2013), we show that the arrival of a star, whom we define as a scientist whose output in terms of citation-weighted publications is above the 90th percentile of the citation-weighted stock of papers published up until year t_{-1} , leads to a significant increase in the productivity of colocated scientists. More specifically, we show this effect operates through two channels: knowledge and recruiting externalities. We show that the arrival of the star leads to an increase in the productivity of incumbents, those scientists already at the department prior to the arrival of the star, but only for those incumbents working on topics related to those of the star. We do not find any evidence of productivity gains by incumbents working in the field of evolutionary biology, but on topics unrelated to those of the star. These effects are robust to including controls for broader departmental and university expansion. Furthermore, they are robust to placebo tests for the timing of the effect; we find no evidence of a pretrend in terms of increasing productivity prior to the arrival of the star. Moreover, the results are also robust to using a plausibly exogenous instrument for star arrival.

The star's arrival also leads to a significant increase in subsequent joiner quality (recruits hired after the arrival of the star), which is most pronounced for related joiners but also occurs for unrelated joiners. These results also hold when subjected to the robustness tests described above. These recruiting results raise a concern about the possibility of reputation-driven positive sorting at top institutions, with stars attracting stars irrespective of productivity-increasing knowledge spillovers. This in turn raises a concern about lost opportunities for stars to seed focused and dynamic research clusters at lower-ranked institutions.

But are these opportunities actually lost? Given the apparent role of star recruitment in department building—which our evidence suggests would be particularly effective where the institution already has a cadre of incumbents working in related areas to the star and has a sufficient flow of new openings to take advantage of star-related recruitment externalities—an offsetting force to excessive concentration could come from the incentive of lower-ranked institutions to use star-focused strategies to ascend departmental rankings. We show how departmental rankings changed between 1980 and 2000 in figure 3.18. While these data imply a reasonably high degree of rank persistence, they also show that some institutions made significant movements up the rankings. Anecdotal evidence suggests that the recruitment of stars may have played an important role here.

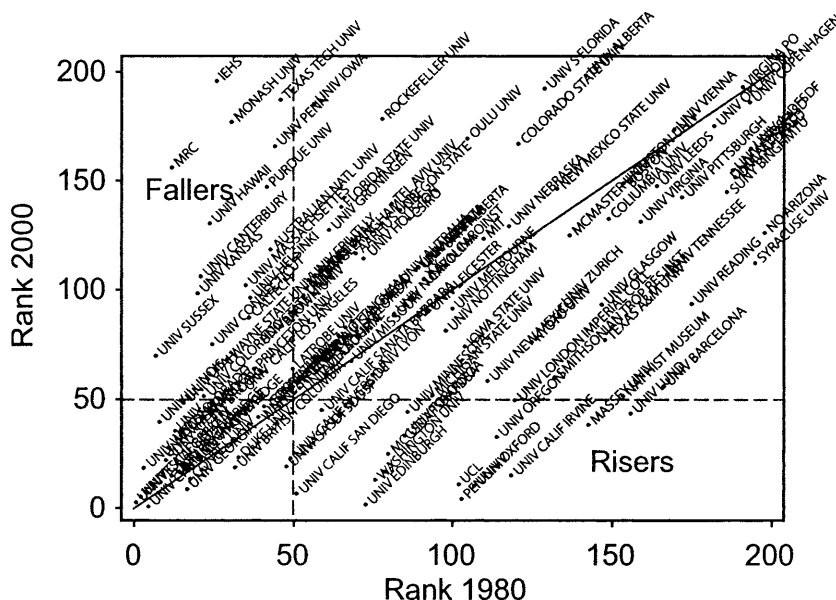
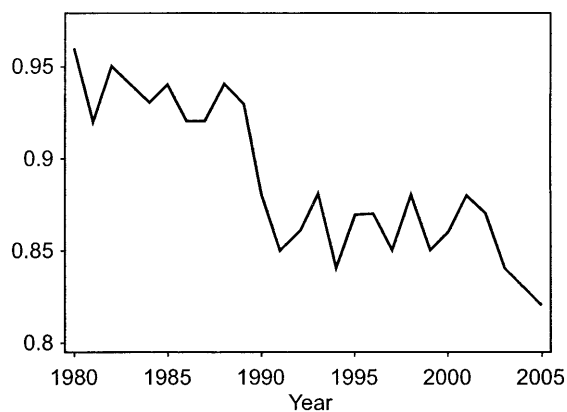


Fig. 3.18 Department-level rank in evolutionary biology: 1980 versus 2000

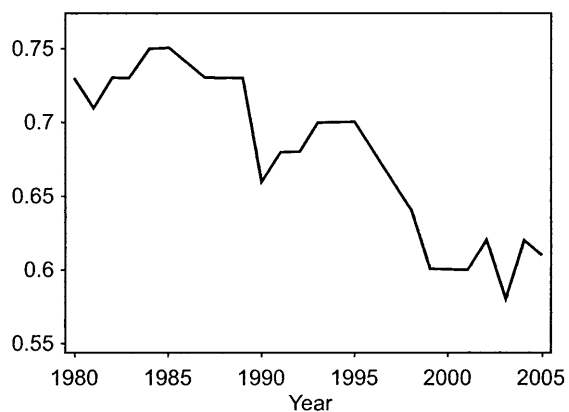
Moreover, stars may increasingly benefit institutions they do not join but where they have collaborative relationships. Azoulay, Graff Zivin, and Wang (2010) and Oettl (2012), who both use the unexpected death of star scientists to estimate their effect on the productivity of their peers, report evidence that stars significantly influence the productivity of their collaborators. Moreover, Agrawal and Goldfarb (2008) show that the greatest effect of universities connecting to Bitnet (an early version of the Internet) in terms of influencing cross-institution collaboration patterns was not between researchers at tier 1 institutions but rather tier 1–tier 2 collaborations. This is consistent with the data we report here on the increasing institution rank distance between collaborators. One interpretation of this result is that lowering communication costs particularly benefits vertical collaboration, suggesting an increasingly vertically disaggregated division of labor as communication costs fall. Perhaps, for example, declining communication costs increase the returns for individuals at top institutions specializing in leading major research initiatives, identifying key research questions, and writing grant applications, while their collaborators at lower-ranked institutions run experiments, collect and analyze data, and work together with all collaborators to interpret and write their results. The results reported by Kim, Morse, and Zingales (2009) are consistent with this when they document the rise of lesser-ranked universities.

To obtain more direct evidence of changes in star concentrations, we plot in figure 3.19 the share of the top 100 evolutionary biology scientists at the

a) Top 100 Scientists in Top 50 Institutions



b) Top 100 Scientists in Top 25 Institutions



c) Top 100 Scientists in Top 10 Institutions

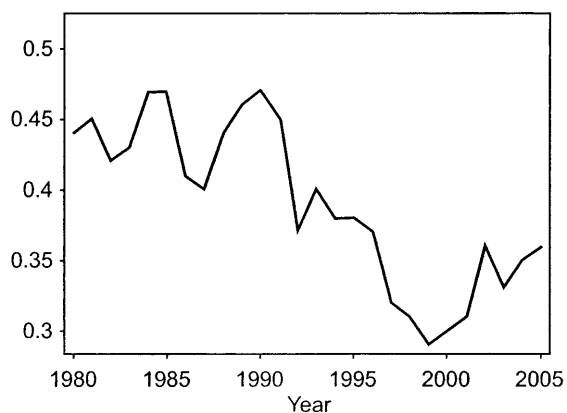


Fig. 3.19 Fraction of top 100-ranked researchers at top-ranked departments

top 50, top 25, and top 10 evolutionary biology departments. The basic pattern shows, if anything, a fall in the concentration of stars at top institutions, somewhat allaying fears of excessive concentration due to reputation-driven positive sorting.

Our examination of the efficiency of the emerging organization of activity in the field of evolutionary biology is unavoidably preliminary and speculative given current levels of knowledge. The broad pattern of increased spatial and cross-institution collaboration—often centered on a star—is pronounced in the data. However, despite institution-level evidence of reputation-based sorting, we do not observe the feared rise in concentration at top institutions. Given the importance of the spatial and institutional distribution of stars to the workings of collaborative science, we expect the normative implications of the changing spatial distribution of scientific activity—and its stars—to be an active area of future research on the organization of science.

References

- Agrawal, A., and A. Goldfarb. 2008. "Restructuring Research: Communication Costs and the Democratization of University Innovation." *American Economic Review* 98 (4): 1578–90.
- Agrawal, A. K., A. Goldfarb, and F. Teodoridis. 2013. "Does Knowledge Accumulation Increase the Returns to Collaboration? Evidence from the Collapse of the Soviet Union." NBER Working Paper no. 19694, Cambridge, MA.
- Agrawal, A. K., J. McHale, and A. Oettl. 2013. "Why Stars Matter." NBER Working Paper no. 20012, Cambridge, MA.
- Azoulay, P., J. Graff Zivin, and J. Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* 125 (2): 549–89.
- Bresnahan, T. F., and S. Greenstein. 1999. "Technological Competition and the Structure of the Computer Industry." *Journal of Industrial Economics* 47 (1): 1–40.
- Gaspar, J., and E. L. Glaeser. 1998. "Information Technology and the Future of Cities." *Journal of Urban Economics* 43 (1): 136–56.
- Jones, B. F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* 76 (1): 283–317.
- Jones, B. F., S. Wuchty, and B. Uzzi. 2008. "Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science." *Science* 322 (5905): 1259–62.
- Jones, C. I. 1995. "R&D-Based Models of Economic Growth." *Journal of Political Economy* 103 (4): 759–84.
- Kim, E. H., A. Morse, and L. Zingales. 2009. "Are Elite Universities Losing Their Competitive Edge?" *Journal of Financial Economics* 93 (3): 353–81.
- Mokyr, J. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.
- Oettl, A. 2012. "Reconceptualizing Stars: Scientist Helpfulness and Peer Performance." *Management Science* 58 (6): 1122–40.
- Ozcan, Y., and S. Greenstein. 2013. "The (De)concentration of Sources of Inventive Ideas: Evidence from ICT Equipment." Working Paper, Northwestern University.

- Romer, P. M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5): S71–102.
- Rosenbloom, R. S., and W. J. Spencer. 1996. *Engines of Innovation: US Industrial Research at the End of an Era*. Boston: Harvard Business Press.
- Tang, L., and J. P. Walsh. 2010. "Bibliometric Fingerprints: Name Disambiguation Based on Approximate Structure Equivalence of Cognitive Maps." *Scientometrics* 84 (3): 763–84.
- Waldinger, F. 2012. "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany." *Review of Economic Studies* 79 (2): 838–61.
- . 2013. "Bombs, Brains, and Science: The Role of Human and Physical Capital for the Creation of Scientific Knowledge." Working Paper, University of Warwick.
- Weitzman, M. L. 1998. "Recombinant Growth." *Quarterly Journal of Economics* 113 (2): 331–60.
- Wuchty, S., B. F. Jones, and B. Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316 (5827): 1036–39.

Comment Julia Lane

The authors address an interesting and important question about the way in which scientific collaboration has changed over time. They use a creatively constructed data set on evolutionary biology to show how scientific collaboration for the subset of authors they identify has changed. The results reported are consistent with other work in the book. Most interestingly, the geographic distance between coauthors has increased substantially, notably that the concentration of publications within an institution has decreased and that the institutional rank distance between coauthors has increased. They find that the concentration of publications at the individual level has increased. They also note the pool of potential coauthors has increased. The authors posit that these trends are the result of two factors: the burden of knowledge and collaboration-supporting technologies.

Their work thus provides potential new areas that could be examined in future research. One is whether evolutionary biology is unique among scientific disciplines: it would be extremely useful to know whether similar changes in collaboration are found for such subsets of authors in "big science," like astrophysics, and smaller scale sciences, like chemistry. It would also be useful to examine across different disciplines whether observed changes in collaboration are due to specialization of innovative labor. It would also be very useful to understand the role of technology in driving geographically dispersed collaboration. It is possible, and anecdotal evidence suggests, that increasingly technology-intensive science that requires large-scale complex equipment is a driving force behind changes in collaboration.

Julia Lane is an Institute Fellow at the American Institutes for Research.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see <http://www.nber.org/chapters/c13039.ack>.

The research also raises interesting questions about whether a field, even though it might be named the same over time, is indeed still the same. Does evolutionary biology still mean the same thing now as it did twenty-nine years ago? Or has it now become more interdisciplinary and been influenced by the convergence of physics, chemistry, and biology? In other words, how much are changes in collaboration due to changes in the very structure of science itself? There is a great deal of interest in using natural language processing techniques to study the evolution of scientific disciplines; future research could examine the role of changes in the topics covered under different disciplinary nomenclatures on collaboration (Herr 2009; Talley et al. 2011).

The findings also suggest interesting potential research into other areas—particularly the role of monetary incentives in changing collaboration. Over the period that is studied, US science agencies changed the way in which science was funded. Although the big interdisciplinary funding initiatives—such as funding for nanotechnology and the human genome sequence—are well known and fundamentally changed both the nature and scale of scientific endeavor in a number of fields, there have been many such smaller scale initiatives. In addition, during the period of study, funding abruptly doubled for the NIH and then leveled off—and so the sharp change in funding could be used to examine the role of monetary incentives.

The authors also posit some interesting hypotheses about the role of graduate students and postdoctoral fellows in the changes in collaborations they observe among their subset of authors. We know that the composition and size of scientific teams is not only dramatically different across research fields, but also changes substantially in response to monetary incentives (Stephan 2007, 2012). Subsequent research on the role of team structure in collaboration could examine both how the structure of teams in evolutionary biology changed over the period—and whether and how subsequent placements of team members in industry or in academia evolved over the time period in question.

Much can also be learned from the extensive hard work that was done to structure the data set. The chapter is partly about authors and partly about institutions, which makes it very difficult to develop a representative frame. The authors have done an enormous amount of work to develop a frame that is based on a very specific selection of publications; this is the link asset that includes information about both.

Future research could be very useful to inform us about the generalizability of the authors' results. In particular, it would be extremely useful to learn whether the known skewness of productivity of scientists, of institutions, and of the salience of publications yields an analysis that is representative of all three dimensions. There has been a rich vein of research in other fields, like labor economics, to understand how the link between workers and firms (jobs) can become a frame of study in its own right (Abowd and Vilhuber 2011; Abowd, Haltiwanger, and Lane 2004; Burgess, Lane, and Stevens 2000);

similar research could usefully be undertaken using data sets constructed in the way described in this chapter. I outline a possible research agenda below.

The Link Asset: Publications. The authors use an imaginative approach to construct a frame from which to draw their sample. In particular, the frame described in this chapter is derived from publications in four journals over twenty-nine years whose focus is evolutionary biology. It would be extremely useful for other researchers to determine whether frames constructed like this are representative of research fields. Are such frames constant over time, both with respect to coverage and content? In particular, how does the emergence of new journals that might siphon off contributors affect the representativeness of the frame? For example, in economics, would a frame based on the publications *Journal of Political Economy*, *Quarterly Journal of Economics*, *American Economic Review*, and *Econometrics*, for example, be equally representative over time, both in levels, and as a proportion of all journals in which economists publish? New research in computer science is beginning to enable us to determine what is missing—by using the same capture/recapture techniques that are used to generate censuses of wildlife, we can determine the coverage of journals over time (Khabsa, Treeratpitu, and Giles 2012; Giles and Khabsa 2014).

The authors use an equally imaginative approach to weighting their sample, and their approach should stimulate an interesting line of research. In particular, they use citations to weight publications, rather than the absolute measure of publications themselves. Returning to my earlier analogy about what we have learned from researching firms, if the focus of the research is firms, then firms can be used as the unit of analysis. If the focus of research is employment, then it is more appropriate to use employment-weighted firms. In the former case, the small start-up has the same weighting as does General Motors. In the latter case, General Motors' behavior will dominate the analysis. Similarly, in this chapter, the results will be dominated by the activities of a few highly productive (and cited) researchers; an important line of analysis for future research would be to examine how the results change when all individual researchers are weighted equally.

The Article Authors. The sample design used in the chapter suggests even more useful research that could be undertaken to build a richer understanding of how the empirical approach can be expanded to understand the activities of active researchers. For example, the original draw from the four journals identifies 171,428 authors; 140,240 of these are dropped because there are no more than two publications linked to their names. Additional research could examine the sensitivity of the results to, for example, distance measures (how many authors are dropped because it is not possible to identify nonunique author names, and understanding the selection bias associated with nonunique Asian author names); time variability (how does the sample selection change over time); and institutional variation (how many institutions are dropped). There are equally interesting questions about the impor-

tance of using fractionally weighted output measures rather than full output measures. In the case of this chapter, the authors decide to weight author output measures by the full publications; since coauthorship has increased overtime, this decision weights output in later time periods more heavily than previous time periods. Much useful research could examine how different weighting choices affect the econometric results. At the least, both weighted and unweighted results should be presented and discussed throughout.

The chapter also identifies an important possible set of sources for collaboration—namely the possibility of faculty coauthoring papers with their postdoctoral fellow and graduate students. This builds on a rich literature that has suggested that within-group specialization is an important feature of modern science (Black and Stephan 2010; Conti, Denas, and Visentin, forthcoming). It would be extraordinarily valuable to build a database that included information about the links between the authors (including the postdoc/graduate student/principal investigator relationship) to establish this hypothesis more generally.

The Institutions. Another important potential line of research that is identified in this chapter is the role of institutional locations, about which so little is known. The structure of the sample identifies some interesting features of collaboration. Just over half (57 percent) of the papers have a single institution listed, and so all authors are located with this institution, and 79 percent of authors are attributed to an institution. Future research could examine how a sample constructed like this might influence our understanding of the role of institutions—particularly focusing on how the structure of institutional affiliations changes over time, both domestically and internationally—and how a selection decision based on institutional affiliations affects the number of coauthors included in the analytical data set.

Finally, the authors put together a very interesting model about the role of graduate and postdoctoral students in generating increasing collaborations. This framework can be used to add to the three alternative hypotheses that are also bruited in the literature. In principle, an empirical exercise could test the model by including data on the pool of available graduate students, data on the role of funding agencies in incentivizing collaboration, the role of technological complementarity, as well as the role of previous collocation.

In sum, this chapter presents some very provocative results, which are very congruent with companion chapters in the book. The challenge to the research community is to extend their interesting results generated from a painstakingly assembled and idiosyncratic sample and determine the generalizability of the approach to both evolutionary biology and other areas of science.

References

- Abowd, J., J. Haltiwanger, and J. Lane. 2004. “Integrated Longitudinal Employee–Employer Data for the United States.” *American Economic Review* 94:224–29.

- Abowd, J. M., and L. Vilhuber. 2011. "National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail." *Journal of Econometrics* 161:82–99.
- Black, G., and P. E. Stephan. 2010. "The Economics of University Science and the Role of Foreign Graduate Students and Postdoctoral Scholars." In *American Universities in a Global Market*, edited by C. Clotfelter, 129–61. Chicago: University of Chicago Press.
- Burgess, S., J. Lane, and D. Stevens. 2000. "Job Flows, Worker Flows, and Churning." *Journal of Labor Economics* 18 (3): 473–502.
- Conti, A., O. Denas, and F. Visentin. Forthcoming. "Knowledge Specialization in PhD Student Groups." *IEEE Transactions on Engineering Management*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2210462.
- Giles, L., and M. Khabsa. 2014. "The Number of Scholarly Documents on the Web." *PLoS One*. doi: 10.1371/journal.pone.0093949.
- Herr, B. W., E. M. Talley, G. A. P. C. Burns, D. Newman, and G. LaRowe. 2009. "The NIH Visual Browser: An Interactive Visualization of Biomedical Research." Information Visualization, 13th Annual Conference, Barcelona, July. Institute of Electronics and Electronic Engineers. http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?tp=&arnumber=5190811&url=http%3A%2F%2Fieeexplore.ieee.org%2Fexpls%2Fabs_all.jsp%3Farnumber%3D5190811.
- Khabsa, M., P. Treeratpitu, and L. Giles. 2012. "AckSeer: A Repository and Search Engine for Automatically Extracted Acknowledgments from Digital Libraries." Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. <http://clgiles.ist.psu.edu/pubs/JCDL2012-AckSeer.pdf>.
- Stephan, P. 2007. "Wrapping It Up in a Person: The Mobility Patterns of New PhDs." In *Innovation Policy and the Economy*, vol. 7, edited by A. Jaffe, J. Lerner, and S. Stern. Cambridge, MA: MIT Press.
- . 2012. *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.
- Talley, Edmund M., David Newman, David Mimno, Bruce W. Herr II, Hanna M. Wallach, Gully Burns, Miriam Leenders, and Andrew McCallum. 2011. "Database of NIH Grants Using Machine-Learned Categories and Graphical Clustering." *Nature Methods* 8:443–44. doi:10.1038/nmeth.1619.