Collaboration, Stars, and the Changing Organization of Science: Evidence from Evolutionary Biology [*]

Ajay Agrawal
University of Toronto and NBER
ajay.agrawal@rotman.utoronto.ca

John McHale
National University of Ireland, Galway
john.mchale@nuigalway.ie

Alexander Oettl
Georgia Institute of Technology
alex.oettl@scheller.gatech.edu

September 2013

Abstract

We report a puzzling pair of facts concerning the organization of science. The concentration of research output is declining at the department level but increasing at the individual level. For example, in evolutionary biology, over the period 1980 to 2000, the fraction of citation-weighted publications produced by the top 20% of departments falls from approximately 75% to 60% but over the same period rises for the top 20% of individual scientists from 70% to 80%. We speculate that this may be due to changing patterns of collaboration, perhaps caused by the rising burden of knowledge and the falling cost of communication, both of which increase the returns to collaboration. Indeed, we report evidence that the propensity to collaborate is rising over time. Furthermore, the nature of collaboration is also changing. For example, the geographic distance as well as the difference in institution rank between collaborators is increasing over time. Moreover, the relative size of the pool of potential distant collaborators for star versus non-star scientists is rising over time. We develop a simple model based on star advantage in terms of the opportunities for collaboration that provides a unified explanation for these facts. Finally, considering the effect of individual location decisions of stars on the overall distribution of human capital, we speculate on the efficiency of the emerging distribution of scientific activity, given the localized externalities generated by stars on the one hand and the increasing returns to distant collaboration on the other.

JEL Classifications: O31, O33, I23, J24, L23.

[*] 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 pre-conference 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.

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 combinatory 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 paper, 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 institutions, 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 towards 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

[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).

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 non-stars, 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 co-location. 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 co-location lowers the cost of collaboration, the rising burden of knowledge increases the returns to co-location. 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 email, 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 co-location, 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 co-location, 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 co-located 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 co-location due to clashing egos and increasing returns to "vertical collaboration" across skill sets located at different institutions. Overall, we find a tendency towards 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 2, we explain the construction of our evolutionary biology data at the institutional and individual levels. In Section 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 4. In Section 5, we present data on the overall rise of collaborative activity and the change in collaboration patterns. In Section 6, we develop a model that offers a potential unified explanation of the facts documented in previous sections. Finally, in Section 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.

2 Data

Our study focuses on the field of evolutionary biology, a sub-field 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.

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 29-year period 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 15,526 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 paper, 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.

[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).

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.

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 50 citations) in both papers. We repeat this process for all papers that list non-unique 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 29-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 3 (by construction), the median is 4, 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).

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% 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% of the 32,955 unique authors to an institution. We drop institutions that do not produce at least one publication in each of the 29 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 29-year period, the minimum number of publications per institution per year is 1 (by construction), the median is 11, the mean is 17.7, and the maximum is 181 (Harvard University in 2005).

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: 1) 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, 2) changing preferences of faculty who have spent more time than their predecessors developing specialized research expertise, and/or 3) 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 towards higher value-added activities as part of the economic development process.

In Figures 1 to 3, we report evidence of the broadening base of science in terms of the declining department-level concentration of scientists, 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 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% of departments produce 60% of all publications in 1980 but only 50% in 2000. Similarly, the top 20% produce 75% of all citation-weighted publications in 1980 but only 60% 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.

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 5. The growth rate of publications from institutions based in BRIC countries (Brazil, Russia, India, China) begins to rise dramatically from the early 1990s onwards 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 US, UK, 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 (ICT) over almost the identical period as our study (1976-2010), Ozcan and Greenstein (2013) examine US patent data and find that although the top 25 firms account for 72% of the entire patent stock and 59% of new patents in 1976, they account for only 55% and 50%, respectively, by 2010. The decline is even more dramatic when they restrict the sample to the ownership of high-quality patents (82% down to 62%). 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" (p.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 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.

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 6, indicate a significant increase in concentration during the 1980s and then relative stability during the following decade. Then, in Figure 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% of scientists produce 70% of output in 1980 but 80% by 2000, with most of the shift occurring in the first decade. Furthermore, in Figure 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 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 co-location, both increase the returns to collaboration. We turn to the topic of collaboration next.

5 Collaboration: Increasing Across Distance and Rank

The trend towards increasing collaboration is a well-documented feature of the changing organization of science (for example, see Wuchty et al., 2007). In fact, the rising role of collaboration is one of the most common themes across the papers in this volume on the changing frontiers of science. For example, Branstetter, Li, and Veloso (2013) state: "Our study suggests that the increase in US patents in China and India are to a great extent driven by MNCs [multi-national corporations] from advanced economies and are *highly dependent on collaborations* with inventors in those advanced economies."[3] Forman, Goldfarb, and Greenstein (2013) explain: "We show that these [geographic distribution of inventive activity] results are *largely driven by patents filed by distant collaborators* rather than by non-collaborative patents or by patents by non-distant collaborators...". Stephan (2013) 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 (2013) 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 (2013) discover through their survey: *"The major factor cited for all types of collaborations was 'unique knowledge, expertise, capabilities.'* Non-co-located 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 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 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.

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

We also observe a dramatic trend in the average rank difference between authors on co-authored papers (Figure 11). For example, in 1980, the average distance in rank between collaborating institutions is approximately 30 (e.g., one collaborator is at an institution ranked number 20 and the coauthor is at an institution ranked number 50); by 2005, the difference increases to approximately 55. Furthermore, we find evidence of increasing distance between collaborators over time. We illustrate this in Figure 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 (2013) present survey evidence indicating that, in general, a large fraction of collaborations occur between scientists who were previously co-located. 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 potential collaborators. For example, stars are likely to have more graduate students and postdoctoral students than non-stars, 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 co-located but now distant potential coauthors. In other words, stars are able to employ this technology over a larger number of previously co-located 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 non-stars. We illustrate this in Figures 13 and 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 non-stars. 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 non-stars.

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 non-stars. So, for example, if the falling cost of communications increases the returns to collaboration such that both a star and a non-star increase their number of collaborations by 1, then the average star may choose the best-suited collaborator from a pool of many previously co-located but now distant potential collaborators, while the non-star can only choose from a pool of few. Even if stars and non-stars 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.

We construct a measure of the size of the pool of previously co-located 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 15, we plot the potential distant coauthor pool size for stars versus non-stars (cumulative number of unique coauthors that were previously co-located 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 15 we only count distant distant coauthors that were previously located whereas in the prior figure there were no distance or prior co-location 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 non-stars. Furthermore, in Figure 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 non-stars. We interpret the ratio as a proxy for the degree of selectivity afforded to stars and non-stars. In other words, a higher ratio for stars versus non-stars indicates that stars collaborate with a smaller fraction of their pool of potential coauthors than non-stars. The figure thus suggests that the relative selectivity of stars versus non-stars 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 non-stars.

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 – i.e., a star concentration effect – and also a broadening institutional base of science.

A key assumption is that relationships with previously co-located 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, Ganguli, and Murciano-Goroff (2013), 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 post docs 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 (email, 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.

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:

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

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:

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

The size of this increase is increasing in n:

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

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: $F(X) = \left(\frac{X}{M}\right)^n$. The density function is then: $f(X) = \frac{n}{X}\left(\frac{X}{M}\right)^n$.

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:

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

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.

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

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:

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

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 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.

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

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

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

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

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 multi-author 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 co-located 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 pre-trend 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 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.

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 19 the share of the top 100 evolutionary biology scientists at the 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, September). 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. University of Toronto mimeo.

Agrawal, A. K., J. McHale, and A. Oettl (2013, July). Why stars matter. University of Toronto mimeo.

Azoulay, P., J. Graff Zivin, and J. Wang (2010). Superstar extinction. *Quarterly Journal of Economics 125*(2), 549–589.

Branstetter, L., G. Li, and F. Veloso (2013). The globalization of r&d: China, india, and the rise of international co-invention. Working Paper.

Bresnahan, T. F. and S. Greenstein (1999). Technological competition and the structure of the computer industry. *The Journal of Industrial Economics 47*(1), 1–40.

Conti, A. and C. Liu (2013). The (changing) knowledge production function: Mit department of biology from 1966-2000. Working Paper.

Forman, C., A. Goldfarb, and S. Greenstein (2013). Information technology and the distribution of inventive activity. Working Paper.

Freeman, R., I. Ganguli, and R. Murciano-Goroff (2013). Why and wherefore of increased scientific collaboration. Working Paper.

Gaspar, J. and E. L. Glaeser (1998). Information technology and the future of cities. *Journal of urban economics 43*(1), 136–156.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The 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), 1259–1262.

Jones, C. I. (1995, August). R & d-based models of economic growth. *Journal of political Economy 103*(4), 759–784.

Kim, E. H., A. Morse, and L. Zingales (2009). Are elite universities losing their competitive edge? *Journal of Financial Economics 93*(3), 353–381.

Mokyr, J. (2002). *The gifts of Athena: Historical origins of the knowledge economy.* Princeton and Oxford: Princeton University Press.

Oettl, A. (2012). Reconceptualizing stars: Scientist helpfulness and peer performance. *Management Science 58*(6), 1122–1140.

Ozcan, Y. and S. Greenstein (2013). The (de)concentration of sources of inventive ideas: Evidence from ict equipment. Northwestern University mimeo.

Romer, P. M. (1990, October). Endogenous technological change. *Journal of Political Economy 98*(5), S71–S102.

Rosenbloom, R. S. and W. J. Spencer (1996). *Engines of innovation: US industrial research at the end of an era.* Harvard Business Press.

Stephan, P. (2013). The endless frontier: Reaping what bush sowed? Working Paper.

Tang, L. and J. P. Walsh (2010). Bibliometric fingerprints: Name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics 84*(3), 763–784.

Waldinger, F. (2012). Peer effects in science: Evidence from the dismissal of scientists in nazi germany. *The Review of Economic Studies 79*(2), 838–861.

Waldinger, F. (2013). Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. University of Warwick mimeo.

Weitzman, M. L. (1998). Recombinant growth. *The Quarterly Journal of Economics 113*(2), 331–360.

Wuchty, S., B. F. Jones, and B. Uzzi (2007). The increasing dominance of teams in production of knowledge. *Science 316*(5827), 1036–1039.

Figure 1: Gini Coefficients by Year for the Distribution of Scientists across Departments

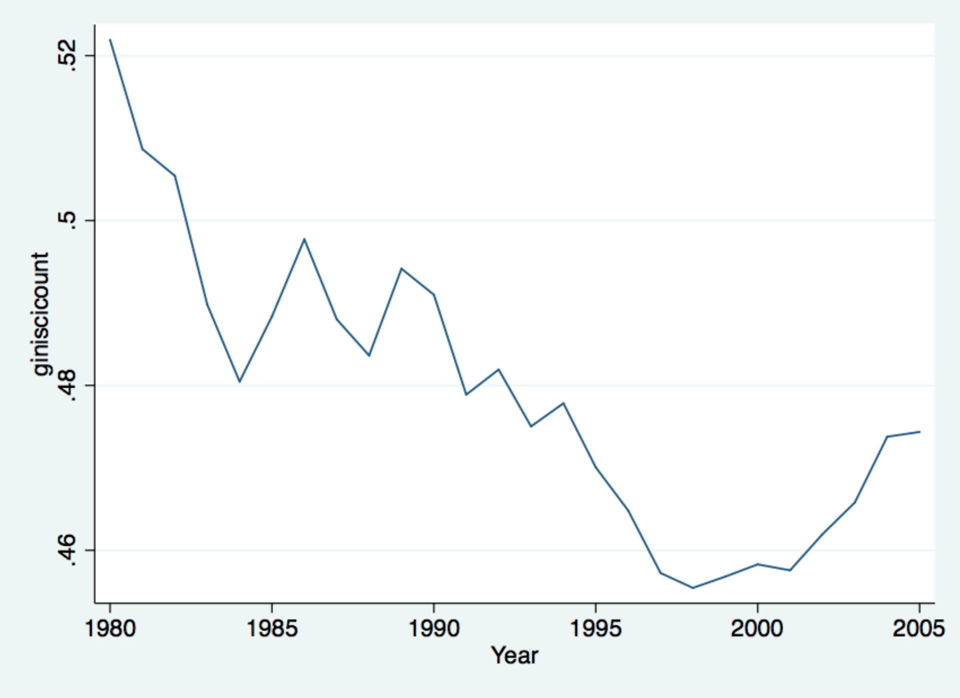


Figure 2: Gini Coefficients by Year for the Distribution of Publications across Departments

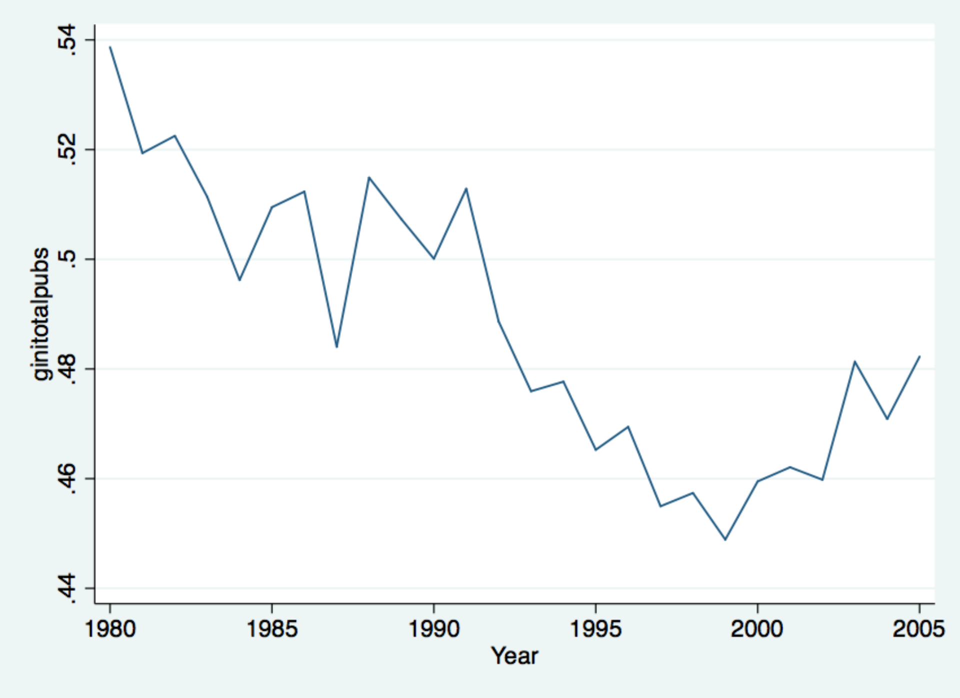


Figure 3: Gini Coefficients by Year for the Distribution of Citations Received across Departments

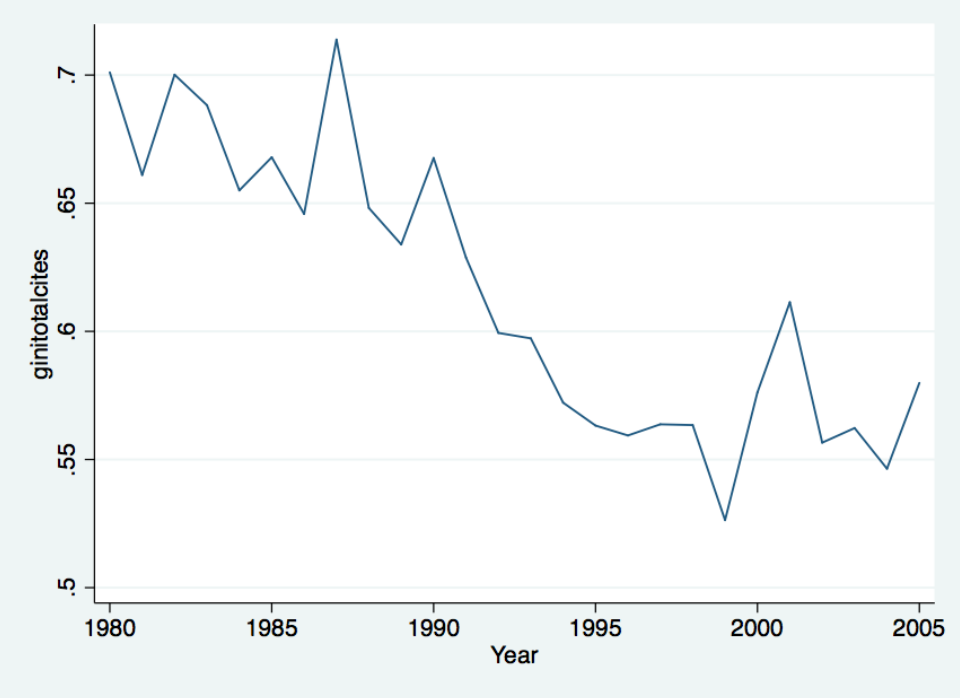


Figure 4: Lorenz Curves by Department for Publications and Citation-Weighted Publications

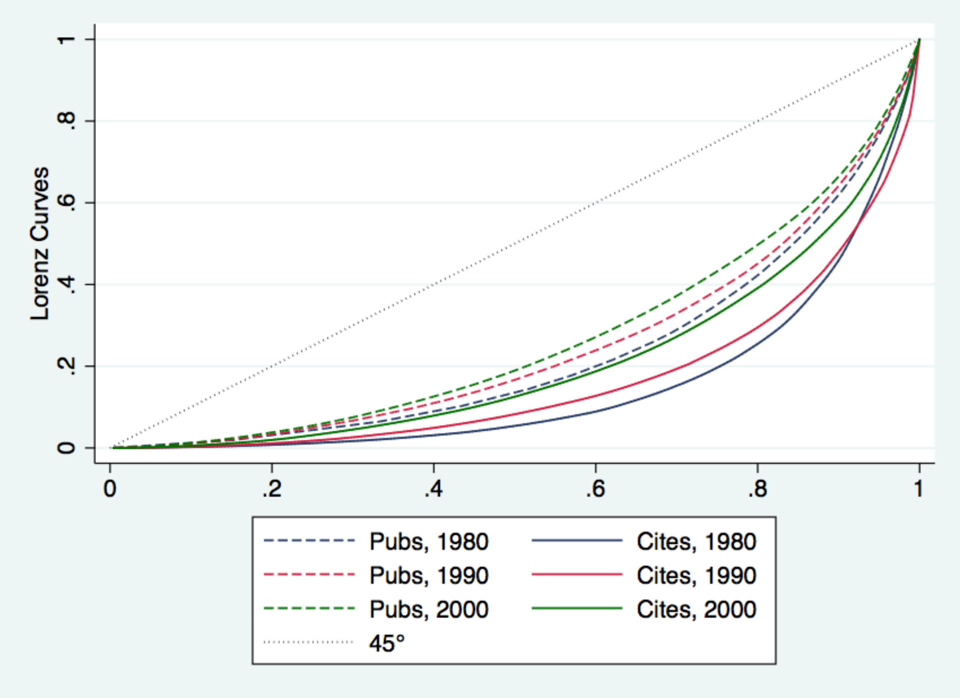


Figure 5: Publication Count by Country by Year Normalized Relative to Output in 1980

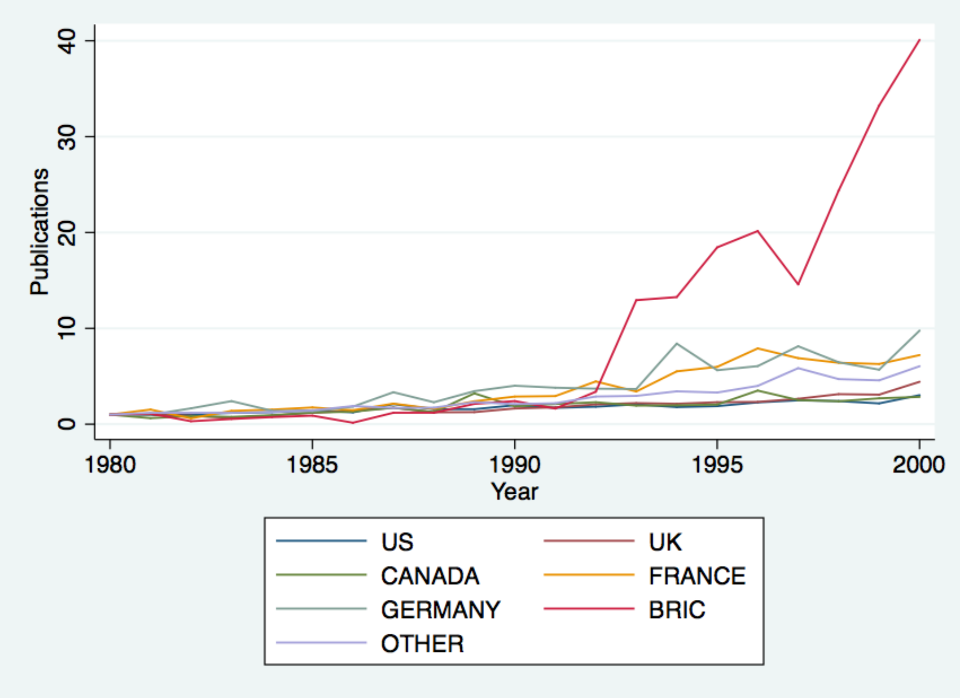


Figure 6: Gini Coefficients for the Distribution of Citation-Weighted Publications across Individuals

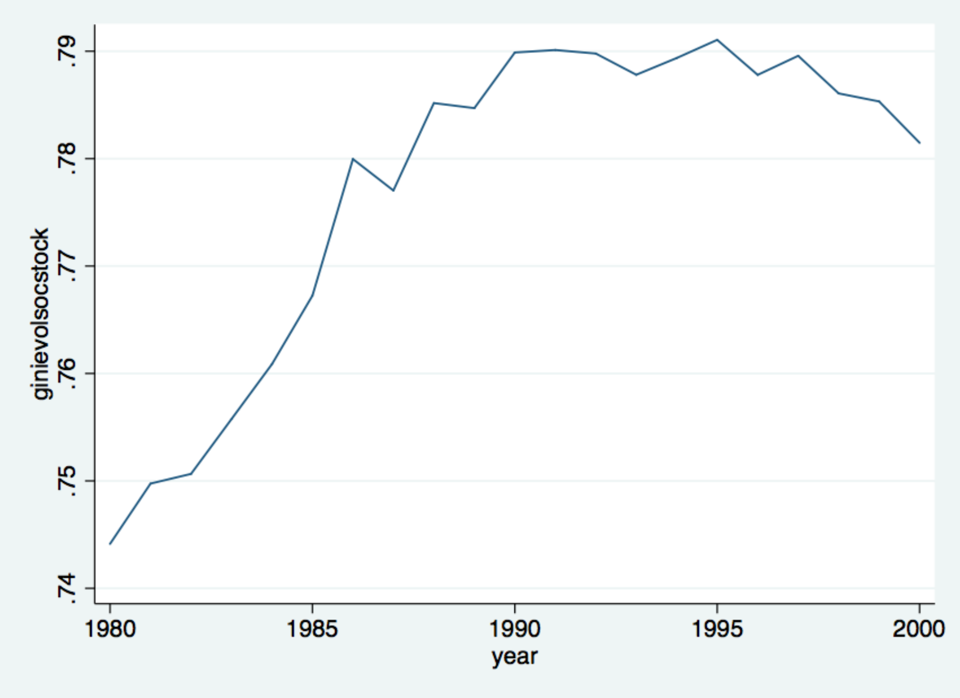


Figure 7: Lorenz Curves by Individual for Citation-Weighted Publications

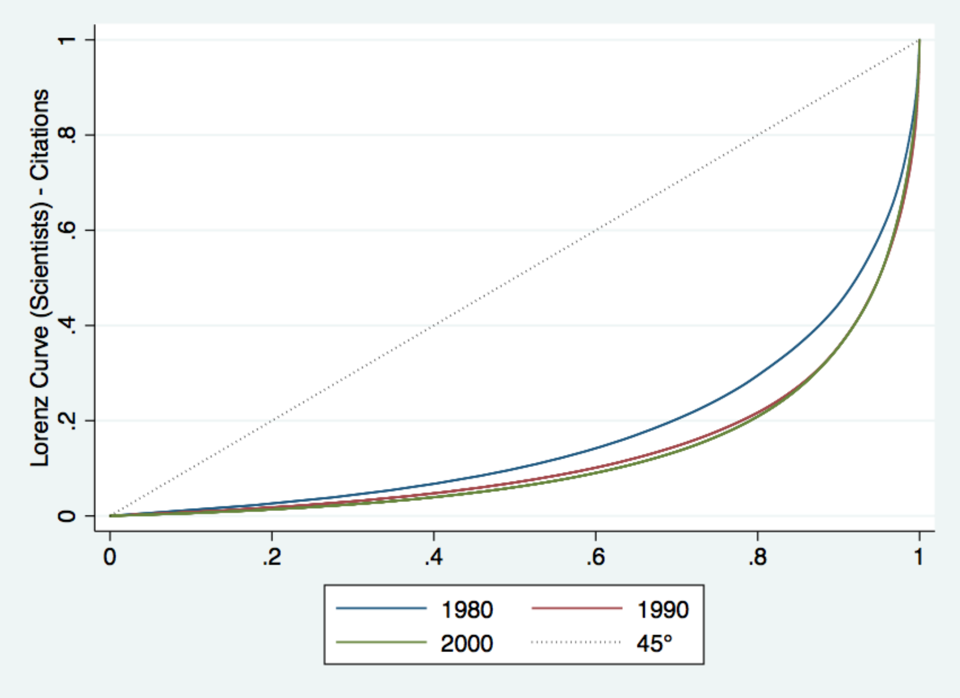


Figure 8: Publication Stock of 50th Ranked Scientist

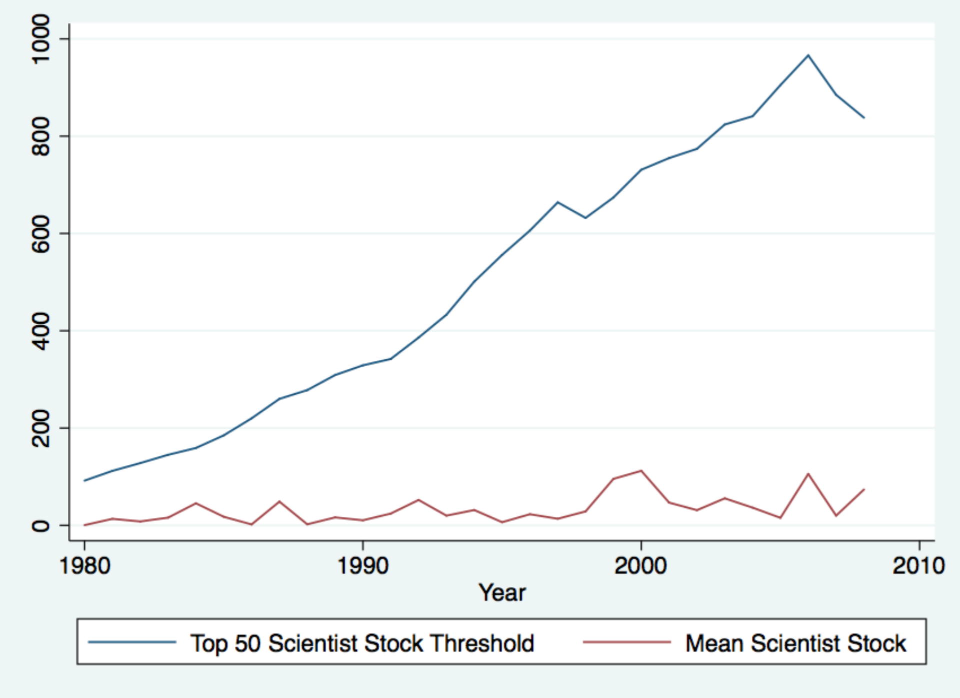


Figure 9: Mean Number of Authors Per Paper

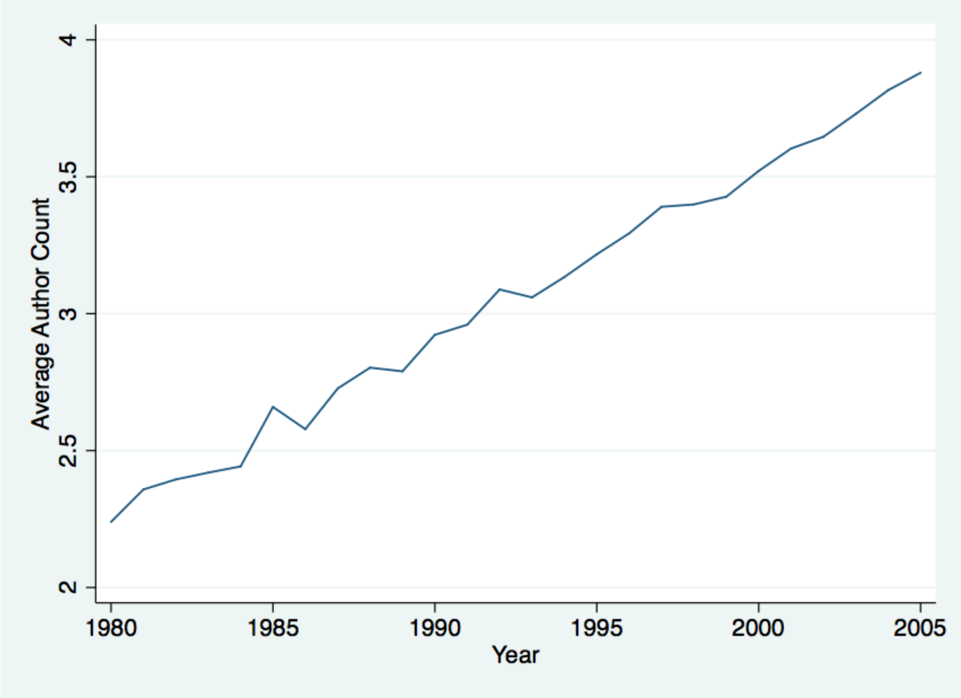


Figure 10: Mean Number of Unique Institutions Per Paper

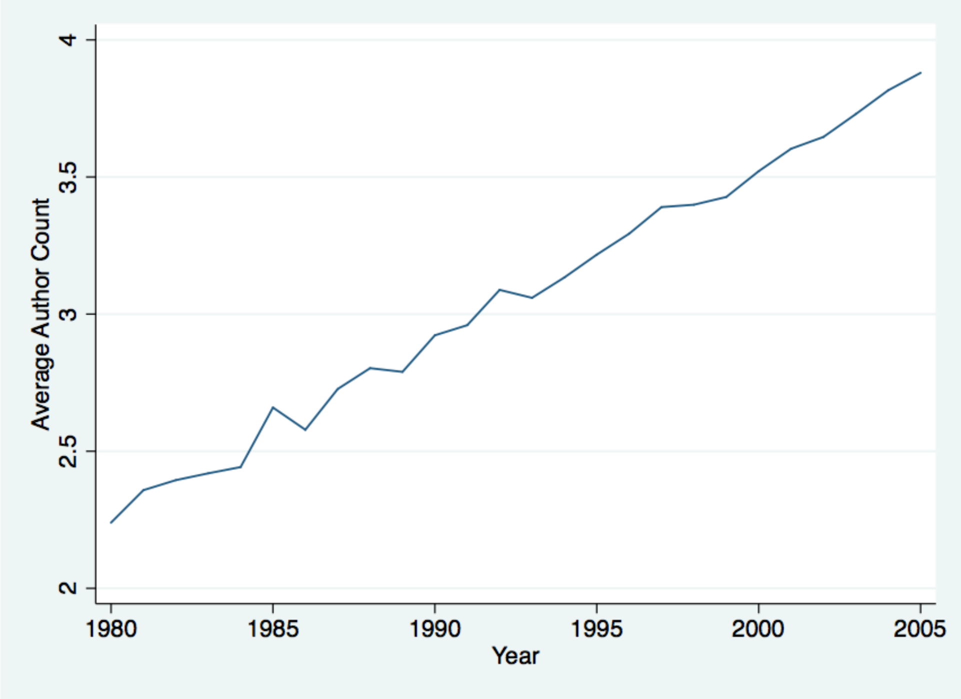


Figure 11: Mean Difference in Institution Rank Between Coauthors

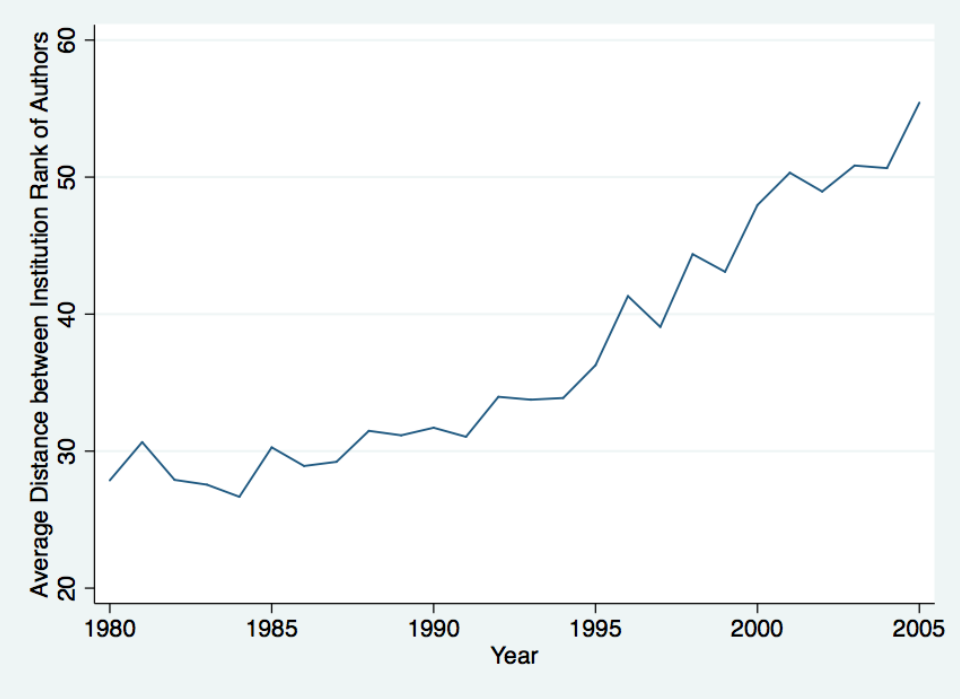


Figure 12: Mean Distance Between Coauthors (miles)

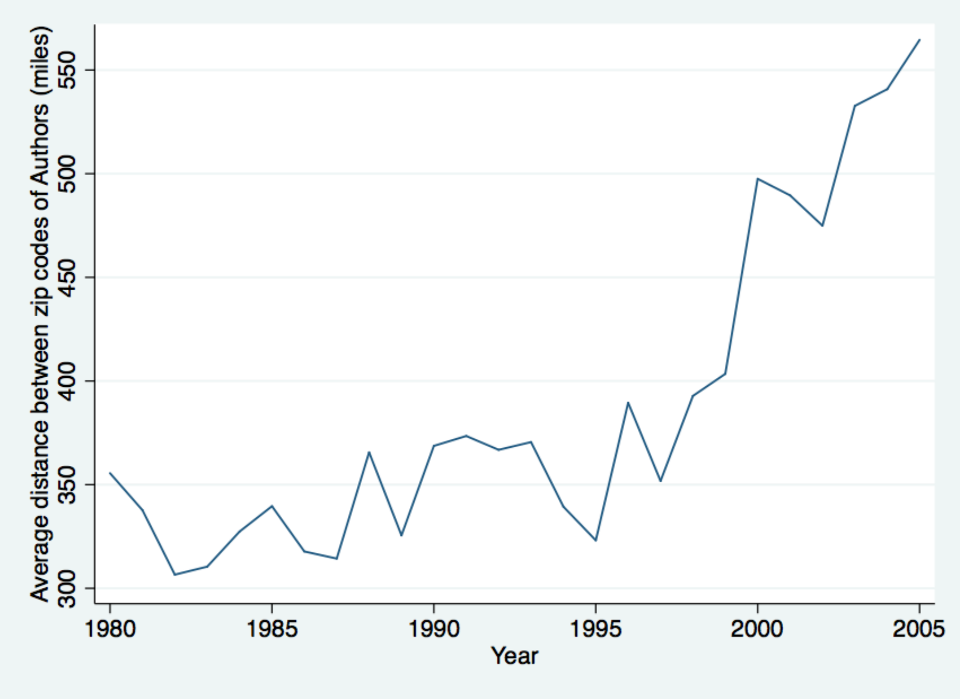


Figure 13: Mean Number of Authors Per Paper (Star versus Non-Star)

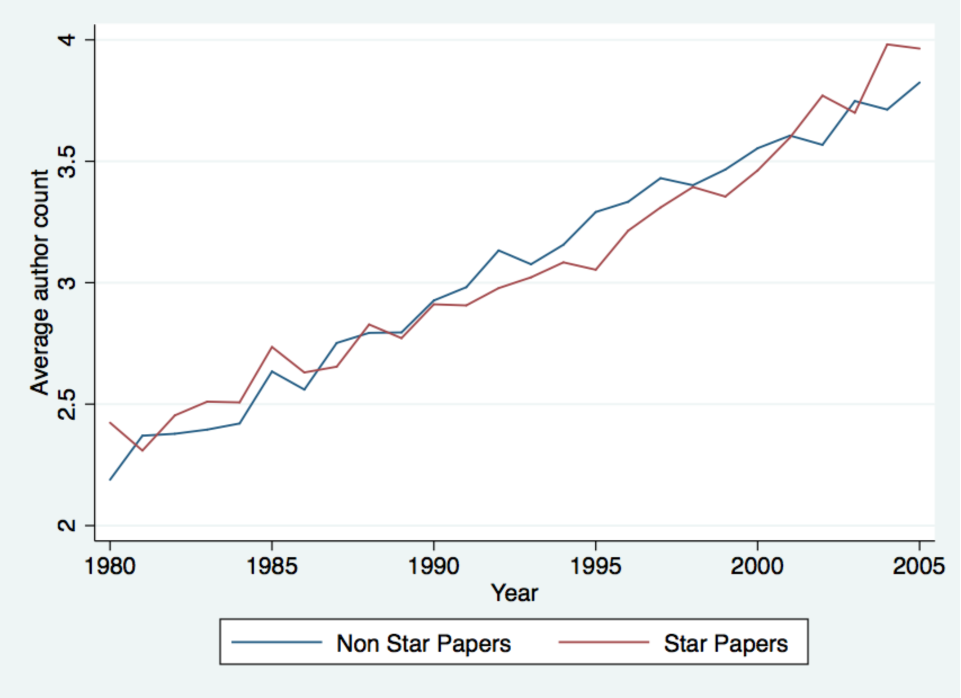
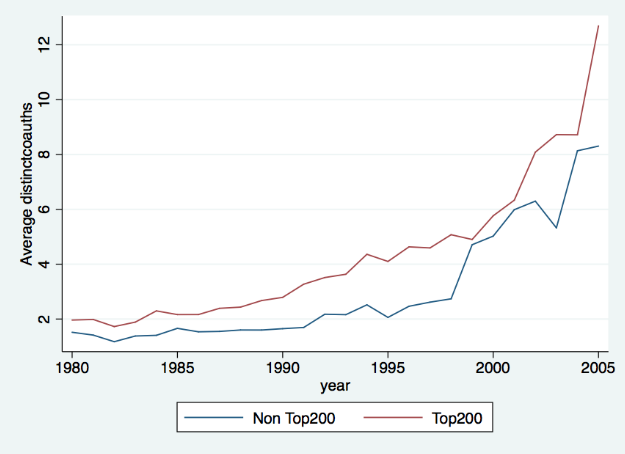
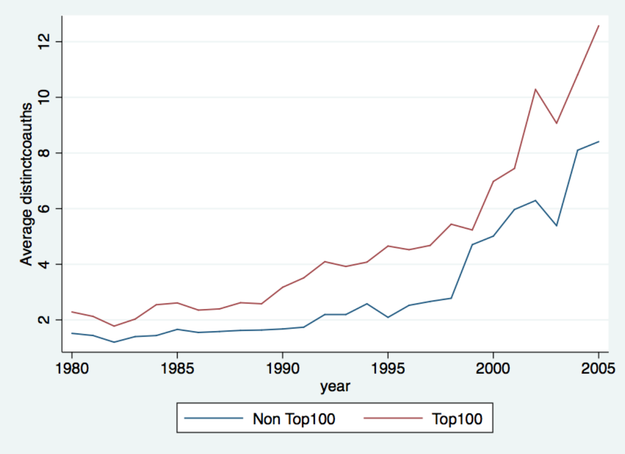


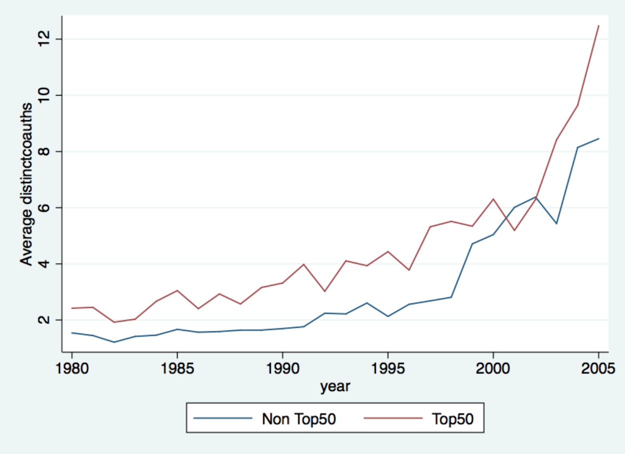
Figure 14: Number of Unique Coauthors Per Year



(a) Top 200 Stars

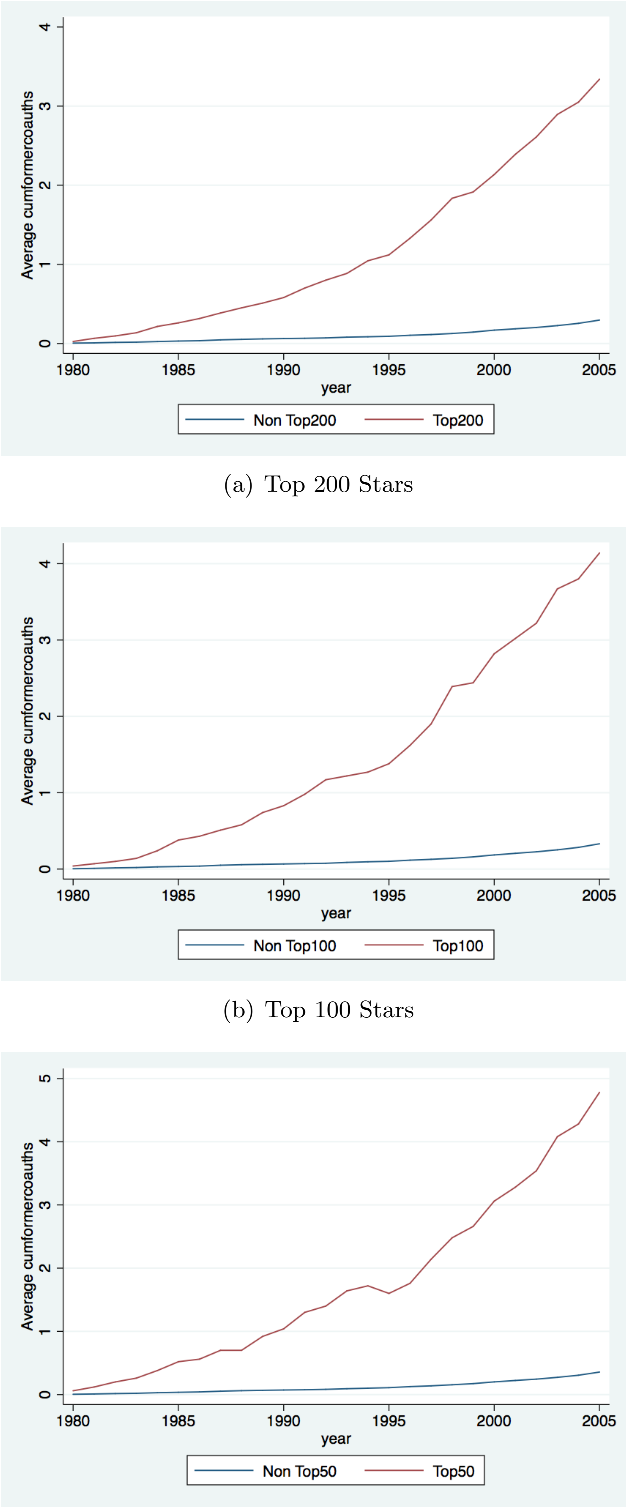


(b) Top 100 Stars



(c) Top 50 Stars

Figure 15: Cumulative Number of Unique Previously Co-Located but Now Distant Coauthors

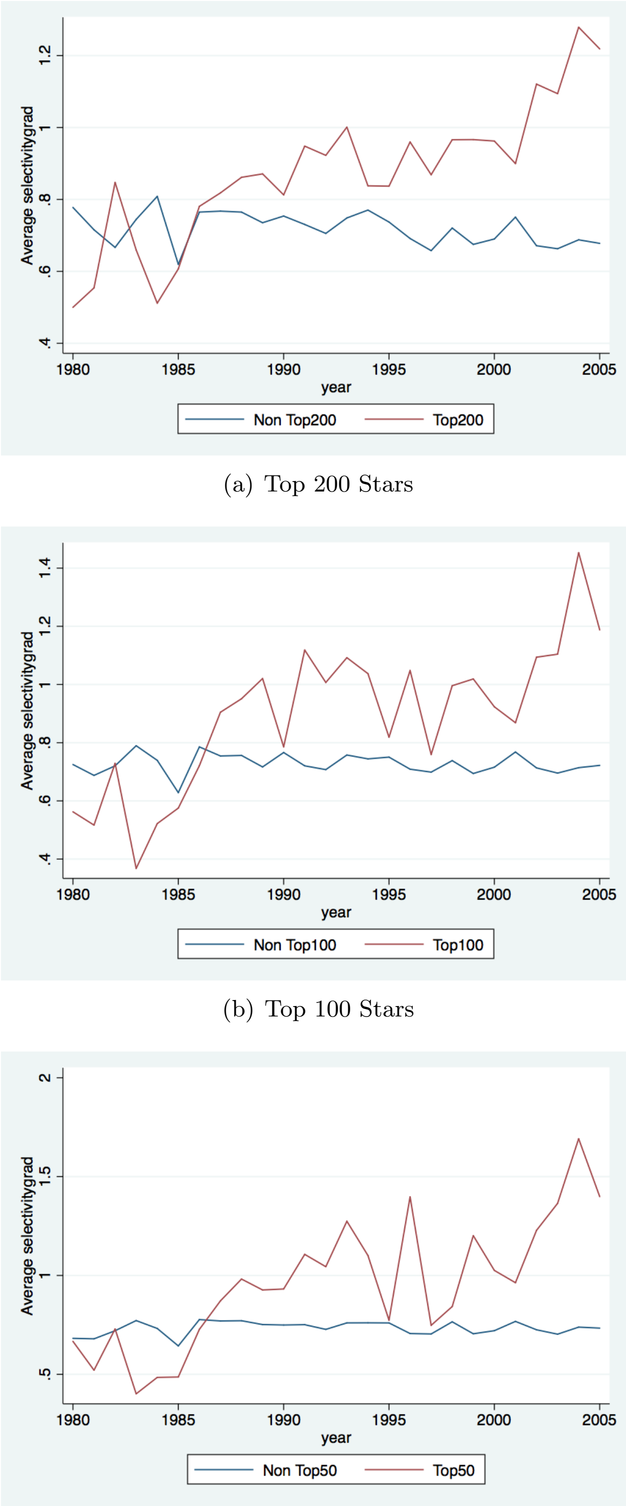


(a) Top 200 Stars

(b) Top 100 Stars

(c) Top 50 Stars

Figure 16: Selectivity Index



(a) Top 200 Stars

(b) Top 100 Stars

(c) Top 50 Stars

Note: We construct the selectivity index as the inverse of the ratio of the number of of unique collaborators in a given year over the cumulative number of previously co-located but now distant collaborators.

Figure 17: Relationship of the Cross Derivative to n

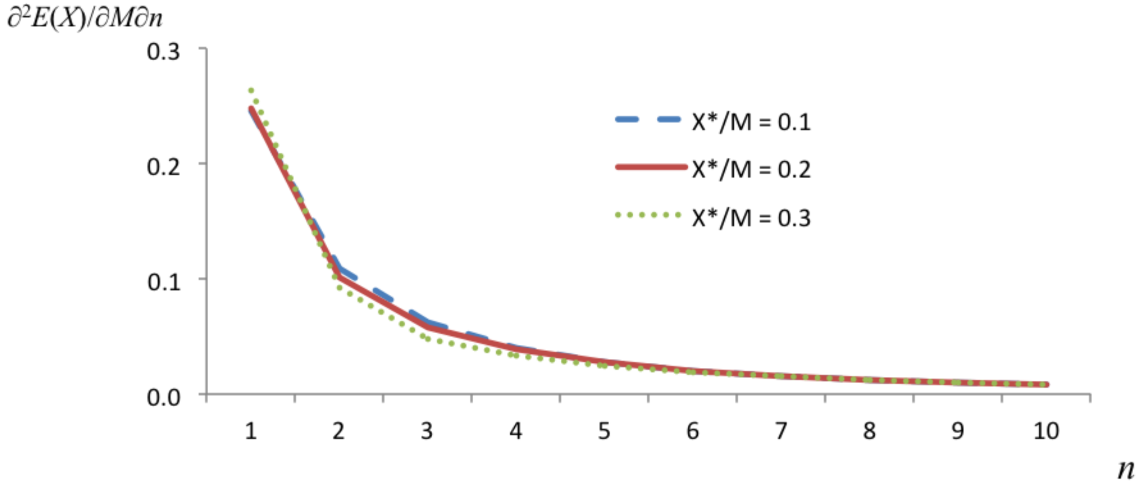


Figure 18: Department Level Rank in Evolutionary Biology: 1980 vs 2000

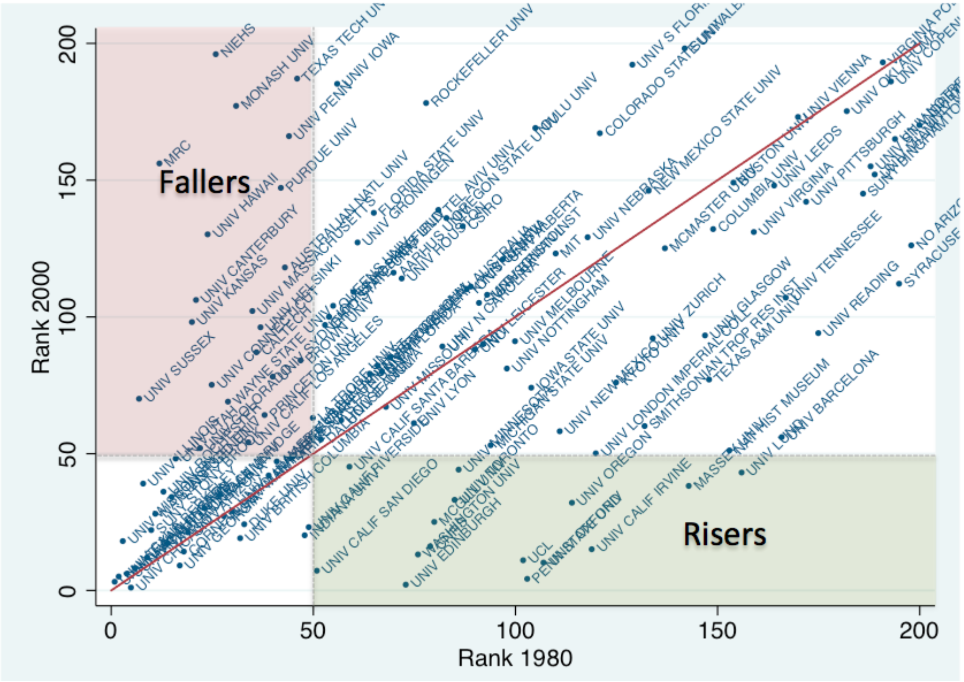
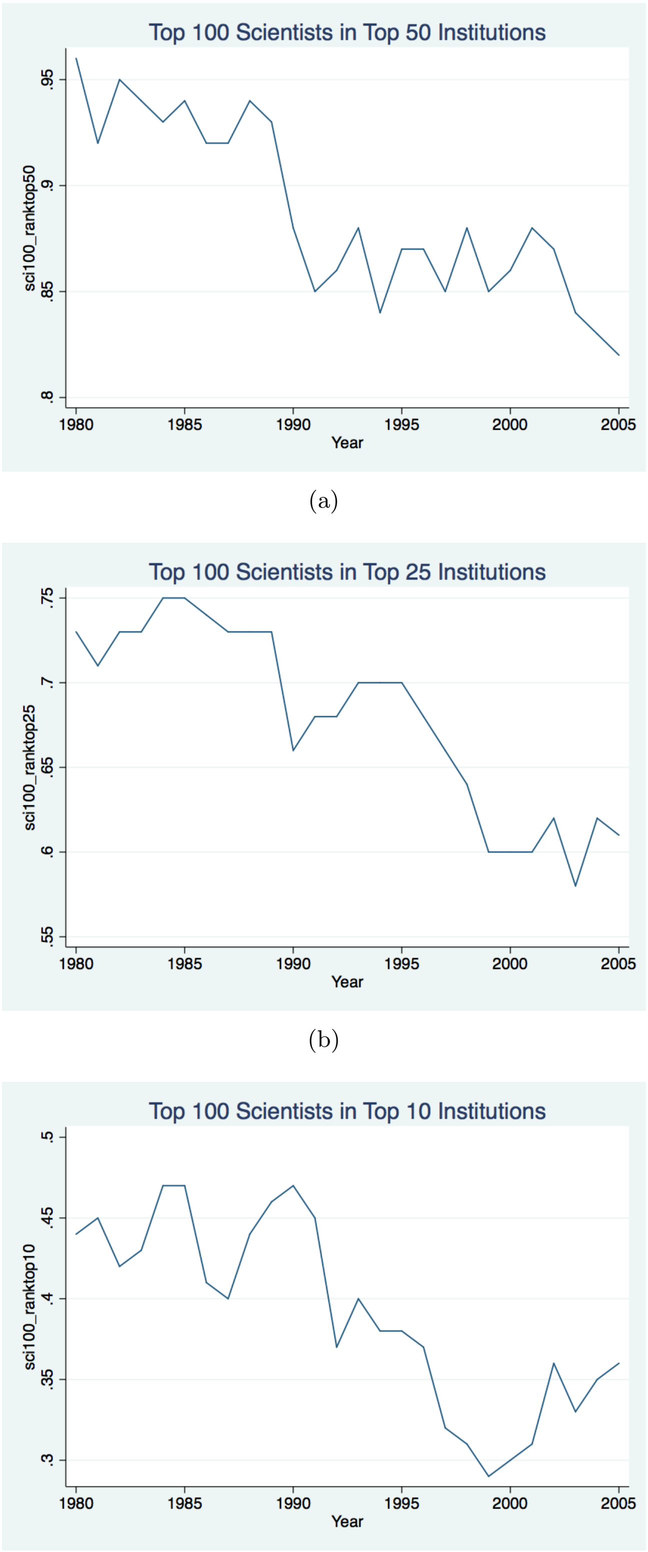


Figure 19: Fraction of Top 100 Ranked Researchers at Top Ranked Departments



(a)

(b)

(c)