

Peer Effects in the Workplace: Evidence from Professional Transitions for the Superstars of Medicine

Pierre Azoulay
Columbia University and NBER
Graduate School of Business
3022 Broadway, 704 Uris Hall
New York, NY 10027

Joshua Graff Zivin
Columbia University and NBER
Mailman School of Public Health
600 West 168th Street, Room 608
New York, NY 10032

November 28, 2005

Abstract

We estimate the magnitude of knowledge spillovers generated by 4,764 academic superstars in the life sciences onto their coauthors' research productivity. Using matched employee-employer data, we measure how scientific output (grants, publications, and patents) for a coauthor changes when the superstar moves to or from a different institution. Preliminary results indicate that superstars generate substantial spillovers through two independent channels: location and coauthorship. Location spillovers decline faster than linearly with geographic distance. Substitution away from collaboration with other scientists cancels a significant portion of the benefits of exposure to superstar talent. We also find that the location spillovers declined markedly in the 1990s.

Keywords: human capital externalities, economics of science, higher education, job mobility, careers.

*Very preliminary and extremely incomplete. Send correspondence to pa2009@columbia.edu or jz126@columbia.edu. Part of the work was performed while the first author was an Alfred P. Sloan Industry Studies Fellow. We gratefully acknowledge the financial support of the Merck Foundation through the Columbia-Stanford Consortium on Medical Innovation. The project would not have been possible without Andrew Stellman's extraordinary programming skills (<http://www.stellman-greene.com/>). For useful comments, we thank Jonah Rockoff.

1 Introduction

Human capital, through its influence on the creation and adoption of improved technologies, has long been recognized as an important contributor to aggregate income (Nelson and Phelps, 1965; Schultz, 1967). Modern economic growth models have built upon this idea to argue that human capital externalities — knowledge spillovers across individuals — are the principal drivers of economic progress (Lucas, 1988; Romer, 1990). The skills and wisdom of some individuals, through interactions with others, increases these attributes in their peers. The increased stock of human capital in the economy generates more ideas and faster growth.

The empirical literature on human capital externalities is relatively small and principally focused on the estimation of peer effects among students (e.g., Rauch, 1993; Acemoglu and Angrist, 2000; Hoxby, 2000; Sacerdote, 2001). Yet, the endogenous growth literature typically envisions spillovers that occur in the workplace, where firms acquire ideas from their neighbors and employees learn from their co-workers. Indeed, the importance of learning through social interactions on the job can be traced back to turn of the century writings by Alfred Marshall (1890). Only a handful of empirical papers have examined peer effects in the employment setting.¹

The paucity of literature is largely due to the difficulties involved in collecting the data necessary to measure these effects. In most employment settings, it is often impossible to identify peer groups at the individual level, and even harder to distinguish output produced jointly with the peer group from output produced independently of it. As a result, the existing research often relies on aggregate proxies — such as co-location — to estimate human capital externalities. Moretti (2004) exemplifies this approach. Using the share of college graduates in the workforce at the city level, he provides evidence for modest human capital externalities in the US manufacturing sector and also finds that spillovers are larger among

¹A related literature examines the influence of peer effects on shirking behavior in the workplace (see for example Ichino and Maggi (2000) or Costa and Khan (2003). Since shirking is easily observed by co-workers and “contagion” does not generally involve the transmission of knowledge or techniques, this type of spillover is conceptually distinct from human capital externalities.

firms that are economically similar. Since manufacturing is less human capital intensive than other sectors of the economy, the finding that spillovers are modest is perhaps not surprising.

In this paper, we attempt to relax the data constraint by focusing on a setting where human capital externalities are likely to be quite important — the academic life sciences. This choice enables us to estimate the productivity benefits of knowledge spillovers at the individual level. Furthermore, it allows us to trace these spillovers back to social interactions, independently of location decisions. More specifically, we analyze the research productivity of coauthors for 4,764 “superstars of medicine” — highly cited researchers, Howard Hughes Medical Investigators, and scientists above the 95th percentile of the NIH award distribution between 1977 and 2003. Using a matched faculty-university panel dataset, we measure how scientific output (grants, publications, and patents) for their coauthors changes when a superstar arrives from — or departs to — a different institution. By limiting our analysis to coauthors and exploiting the longitudinal structure of the data, we are able to avoid some of the traditional concerns regarding endogenous peer selection (Manski, 1993). Moreover focusing on each coauthor-superstar pair permits the separate identification of those spillovers that stem from co-location from those that stem from co-authorship *per se*.

Looking at *changes* induced by location choices could introduce other sources of bias, however. Professional transitions are not exogenous, and if movement between institutions is mostly driven by expectations about the productivity of the new (or old) departments, then estimates of the spillover effect will be biased. While the current version of the paper ignores endogeneity, in future versions we plan to adopt a number of different approaches to address this issue. First, we might focus on 139 superstars who died relatively suddenly while at the peak of their careers. Second, we can make use of information about children’s age and spousal occupation for the superstars to provide plausibly exogenous variation in the propensity to change academic affiliations. Third, we can propose a model of selection on observables in which features of the “peer environment” (number of coauthors, funding trends for the origin institution, etc.) figure prominently among the determinants of job mobility for the superstar.

Our preliminary results are suggestive of the role of exposure to superstar talent in the production of scientific knowledge. We estimate location spillovers that decline quickly with geographic distance, and returns to co-authorship that are both economically meaningful and statistically significant. There is also a clear pattern of substitution whereby output coauthored with others decreases significantly when superstar and colleagues are co-located, but not to the extent of completely undoing the external benefits of collaboration. Finally, we find that the spillovers attributable to co-location declined precipitously in the 1990s.

The rest of the paper proceeds as follows. In the next Section, we present a simple theoretical framework. Section 3 presents our sample of superstars, describes the matching process necessary for the identification of peers. Section 4 describes our presents descriptive statistics, and reports our econometric results. Section 5 concludes.

2 Theoretical Framework (very loose at this point)

Knowledge can flow across time and space through three distinct mechanisms. The endogenous growth literature emphasizes the “standing on shoulders of giants” effect, whereby the current state of knowledge forms the basis for new knowledge-based innovation. This implies a role for education and training, which enable would-be innovators to be active at the frontier of knowledge (Jones, 2005). A second mechanism for knowledge diffusion is *involuntary* spillovers, whereby innovators obtain knowledge generated by others through observation and imitation. Finally, knowledge can be shared directly among members of innovating teams. In this paper, we are interested in evaluating the extent to which such collaborations can foster the production of new knowledge.

Looking at the effect of collaborations is important because very few innovations result from the efforts of lone researchers engaged in otherworldly contemplation. The creative process is characterized by the search for useful recombinations of existing ideas (Weitzman, 1998), and the mixing of individuals with different backgrounds and education within the same team magnifies the number of combinations that can be evaluated. Moreover, members of these teams do not have to be co-located. In fact, they are increasingly distributed across

different locales, thanks to the use of information and communication technologies (Kim et al., 2005; Rosenblat and Mobius, 2004).

When members of different skill level or experience match, less skilled agents might gain from their exposure to the ideas of the relatively more skilled, thus creating *voluntary* spillovers of knowledge. A skilled agent might match with a less skilled one by pure altruism, because of cultural norms that favor mentorship, or because the costs of collaboration with less-skilled agents are lower. Here, we will take the existence of such heterogeneous matches as given, and study their impact on the output produced by the less skilled agents. The impact of the skilled on the less skilled is especially interesting because the former might find it difficult to fully appropriate the benefits they confer on the latter. For example, it might be impossible to apportion credit between team members in ways that are verifiable by a court. Moreover, even when individual talent is perfectly observed in the market, asymmetric information over the mentorship abilities of skilled agents might both limit their mobility and ensure that they cannot extract their full marginal product (Lazear, 1986; Acemoglu and Pischke, 1999).

That knowledge spillovers are at least partially external means innovators will have too little incentive to innovate (Murphy et al., 1991). In such a world, the allocation of talent across firms and the technologies and policies that influence the flow of information between agents has important implications for the level and rate of technological innovation within the economy, and eventually for economic growth.

The setting chosen for our empirical work is the academic life sciences. This sector is an important one to study for several reasons. First, technological change has been enormously important in the growth of the health care economy, which accounts for roughly 15% of US GDP. Much biomedical innovation is science-based (Henderson et al., 1999), and interactions between academic researchers and their counterparts in industry appear to be an important determinant of research productivity in the pharmaceutical industry (Cockburn and Henderson, 1998; Zucker et al., 1998). Second, and perhaps most importantly for our work, academic scientists are generally paid through soft money contracts. Salaries depend on

the amount of grant revenue raised by faculty, thus providing researchers with high-powered incentives to remain productive even after they secure a tenured position. As such, academic life scientists can be viewed as entrepreneurs producing new knowledge through the management of small idiosyncratic firms, comprised of junior faculty, postdoctoral researchers, graduate students, and other research personnel. Lastly, there are large public subsidies for biomedical research in the United States. With an annual budget of \$28 billion in 2004, support for the NIH dwarfs that of other national funding agencies in developed countries (Cech, 2005). Thus, estimating knowledge spillovers in this sector will allow us to better assess the return to these public investments.

Our focus on this setting also offers three practical benefits that better enable us to examine the magnitude of human capital externalities. First, by matching data on research output (grants, publications, and patents) with an administrative dataset linking medical school scientists with their employers, we are able to create what is, to our knowledge, the first matched employee-employer dataset with individual-level measures of output.

Second, the setting affords us the ability to avoid using co-location as a *de facto* measure of social interaction. In principle, there exist three dimensions along which individuals engaged in the production of abstract knowledge could be said to be “close” or “distant.” The first dimension is that of geography; it is the one most often used in the literature (Jaffe et al., 1993). The second dimension is that of scientific or intellectual distance — the extent to which the skills of researchers overlap or complement one another.² Many science policy makers hold the belief that collaborations spanning discipline boundaries are more likely to result in scientific breakthroughs (Metzger and Zare, 1999). Whether the benefits of such “distant” collaborations are large enough to outweigh the coordination costs they entail is an empirical question that deserves closer scrutiny (Jones, 2005).

The third dimension is that of social distance, which can be computed once every scientist is viewed as a node in a network of collaborations defined by co-authorship or co-inventorship links (Fafchamps et al., 2004; Mairesse and Turner, 2005). A large literature in the soci-

²Jaffe (1986) and Adams and Jaffe (1996) apply this concept of distance to firms.

ology of science emphasizes the role of these connections in fostering academic professional achievement. Eminent scientists might not only be a source of insights or funding, but also “annoint” their junior collaborators and enhance their visibility in the scientific community. Our study is the first in which these three concepts of distance can be measured concurrently, and their influence on research output examined simultaneously.

A final advantage of studying the academic labor market in the life science has to do with the frictions created by high mobility costs. Academic medical researchers are frequently married to other academics and “power couples” tend to be less mobile (Sobecks et al., 1999; Costa and Khan, 2000). Star scientists often have large laboratories and it is difficult to make arrangements to move junior faculty, postdocs, and other lab employees to new institutions. For scientists with clinical practices, established clientèle and state-specific physician licensure requirements further hinder movement. These switching costs create at least a presumption that knowledge spillovers from superstars will not be fully internalized by the labor market. This provides additional motivation for studying the impact of superstars on the allocation and development of talent within this industry.

3 Data and Sample Characteristics

This section provides a detailed description of the process through which the matched coauthor/superstar data used in the econometric analysis was assembled. In order, we describe (1) the criteria used to select our sample of superstar life scientists; (2) the universe of potential peers for these superstars; and (3) the essence of the matching procedure implemented to identify actual peers from coauthorship records. We also present basic demographic characteristics for the superstars, as well as descriptive statistics for the individual coauthors and superstar/coauthor dyads.

3.1 Superstar Sample

Our basic approach is to rely on the mobility of “superstar” scientists across academic institutions to estimate the magnitude of knowledge spillovers onto colleagues. This focus on

superstars can be justified on both substantive and pragmatic grounds. Significant inequality in scientists' productivity has been widely documented. In a classic paper, Lotka (1926) showed that the most productive 6% of publishing physicists produced 50% of the papers in the journals he examined. An extensive literature in the sociology of science presents further evidence of the skewed distribution of productivity (e.g., Merton, 1973; de Solla Price, 1986). In a related vein, Zucker et al. (1998) established a robust correlation between the location of superstar life scientists and the number of new biotechnology firms spawned in a given locale. Thus, if one wants to find evidence of spillovers at the individual level, it appears to logical to start with a sample of superstars, rather than with a random sample of scientists. From a practical standpoint, it is more feasible — though still surprisingly difficult — to trace back the careers of eminent scientists than to perform a similar exercise for less eminent ones. We rely on three different measures of scientific achievement to draw a list of 4,764 superstar scientists.

- **Cumulative NIH Funding.** Our first source is the Compound Grant/Applicant File (CGAF) from the U.S. National Institutes of Health (NIH). This dataset records information about grants awarded to extramural researchers funded by the NIH since 1938. Using the CGAF, and focusing only on research grants, we compute individual cumulative totals for the years 1977 to 2003, deflating the earlier years by the biomedical research producer price index. We also recompute these totals excluding large center grants that usually fund groups of investigators (M01 and P01 grants). Those scientists whose totals lie in the top ventile (i.e., above the 95th percentile) of either of these two distributions constitute our first group of superstars. In this elite group, the least well-funded investigator had garnered \$10.5 million in NIH career funding, and the most well-funded \$462.6 million.
- **Highly Cited Scientists.** Despite the preeminent role of the NIH in the funding of public biomedical research (Cech, 2005), this indicator of “superstardom” biases the sample towards older scientists conducting relatively expensive research. We complement this first group with a second composed of highly cited scientists identified by the

Institute for Scientific Information. A Highly Cited listing means that an individual was among the 250 most cited researchers for their published articles between 1981 and 1999, within a specific scientific field.³

- **Howard Hughes Medical Investigators.** Finally, we add to these two groups medical school faculty who are also Howard Hughes Medical Investigators. Every three years, the Howard Hughes Medical Institute solicits nominations from research institutions, with the aim of identifying researchers who have the potential to make significant contributions to science. Once selected, they continue to be based at their institutions, typically leading a research group of 10 to 25 students, postdoctoral associates and technicians. From our point of view, HHMIs are attractive in that they tend to be younger, “up-and-coming” scientists, rather than established investigators.

Table 1 presents the distribution of superstars by achievement criteria. Many scientists achieve superstar status according to more than one metric. We trace back these scientists’ careers with great care from the time they obtain their first position as independent investigators (typically after a postdoctoral fellowship) until 2005. We do so through individual CVs, *who’s who* profiles, accolades/obituaries in medical journals, and google searches. As a result, we will be able to follow their path from humble beginnings to superstardom, and to examine whether spillovers vary over the life cycle, even though selection into the sample was based on *cumulative* achievement.

We record employment history, degree held, date of degree, gender, up to three departments, country of birth, and whether the star held an administrative position, such as dean or hospital CEO. We cross-reference the superstar sample with other measures of scientific eminence: election to the National Academy of Sciences or the Institute of Medicine; Nobel laureates; and winners of the Lasker award. We also match the sample with patent data from the USPTO. 1,752 superstars (37.8% of the sample) are listed as an inventor on at least one patent (to fix ideas, Azoulay et al. (2005) finds that the proportion in a random sample of life scientists is about 16%), and 127 hold 17 patents or more, which puts them above

³The relevant scientific fields in the life sciences are microbiology; biochemistry; psychiatry/psychology; neuroscience; molecular biology & genetics; immunology; pharmacology; and clinical medicine.

the 99th percentile of the cumulative patent distribution (where the universe is restricted to patents with at least one academic or hospital assignee).

Table 2 provides descriptive statistics for the superstar sample. The sample is 11.84% female and 81.38% US-born. 45% of our stars hold an MD degree, 43% a PhD, and the remainder hold dual MD/PhD degrees. The oldest graduated in 1927, the youngest in 1998, while the modal scientist in the sample graduated in 1970. Table 3 lists the Top 20 research institutions among those that employed our superstars. Unsurprisingly, the list bears a close resemblance with that of top recipients of NIH grants. Table 4 presents information about the superstars' main department affiliation at the end of career or in 2003, whichever comes earlier. The sample is approximately evenly divided between clinical and "basic" scientists. Table 5 reports on job mobility patterns among superstars. 45% of these scientists never move, which is not too surprising, given the transaction costs involved in moving a laboratory to another institution. In practice, these sedentary scientists will still be useful in the estimation of peer effects since data on their colleagues will help pin down age, cohort and calendar year effects. Some of the superstars exit the sample in or before 2005: 574 (12.05%) retire; 139 (2.92%) die while still holding a faculty position; and slightly less than 2% exit to industry, government positions, or foundations.

Beyond these demographic characteristics, we can describe productivity patterns in the superstar sample along three distinct dimensions: NIH grants, publications, and patents. To adjust publications for quality, we make use of the Journal Citation Reports, published yearly by the Institute for Scientific Information. ISI ranks journals by impact factor (JIF) in different scientific fields. The impact factor is a measure of the frequency with which the "average article" in a journal has been cited in a particular year. We weight each article published by the scientists in our sample by the corresponding journal's JIF, and compute quality-weighted publication counts in this way.⁴

⁴Basically a ratio between citations and recent citable items published, JIFs suffer from built-in biases: they tend to discount the advantage of large journals over small ones, of frequently-issued journals over less frequently-issued ones, and of older journals over newer ones. Nonetheless, they convey quite effectively the idea that the *New England Journal of Medicine* (Impact Factor = 23.223 in 1991) is a much more influential publication than the *Journal of General Internal Medicine* (Impact Factor = 1.056 in 1991).

Figures 1, 2 and 3 display histograms for the distributions of cumulative career NIH funding, cumulative publication counts, and cumulative patents, respectively. Even in this selected group, the extent of heterogeneity in output is striking. Of course, not all of this variation can be attributed to talent. A large part undoubtedly reflects life cycle and cohort influences (Levin and Stephan, 1991).

3.2 The Universe of Potential Peers

Information about superstars' peers stems from the faculty roster of the American Association of Medical Colleges, which we secured for the years 1977 through 2003 under a confidentiality agreement. The roster is an annual census of all U.S. medical school faculty, where each faculty is linked across yearly cross-sections by a unique identifier. When all cross-sections are pooled, we obtain a matched employee/employer panel dataset. For each of the 202,292 faculty members that appear in the roster, we know the full name, the type of degrees received and the years they were awarded, gender, up to two departments, and medical school affiliation. Academic rank information is also available, but it is not recorded in each year. Besides its comprehensiveness, an attractive feature of this data source is that it shares a common system of individual identifiers with the CGAF dataset.

Because the roster only lists medical school faculty, however, it is not a complete census of the academic life sciences. For instance, it does not list information for faculty at institutions such as MIT, UC-Berkeley, Rockefeller University, the Salk Institute, or the Bethesda campus of the NIH; and it also ignores faculty members in Arts and Sciences departments — such as biology and chemistry — if they do not hold joint appointments at a local medical school.⁵

To identify peers, we focus on those faculty listed on the roster that are coauthors for each superstar. Although one could identify peers according to different criteria, upon reflection a coauthorship-based definition seems to us the most sensible. Specifically, we could have

⁵This limitation is less important than might appear at first glance. First, we have no reason to think that colleagues located in these institutions differ in substantive ways from those based in medical schools. Second, all our analyses focus on *changes* in research productivity over time for a given scientist. Therefore, the limited coverage is an issue solely for the small number of faculty who transition in and out of medical schools from (or to) other types of research employment. For these faculty, we were quite successful in filling out career gaps by combining the roster with the NIH data.

labeled peers any faculty that was co-located with the superstar at any point during his or her career. This does not strike us as a meaningful definition of the term “peer,” because medical schools are large research institutions (833 faculty on average in 2003). One could make use of department information to define narrower boundaries for peer groups, but this approach is too difficult to implement in practice. First, department affiliations are not fixed over time for most faculty — this is apparent in our sample of superstars, and many collaborations span departmental boundaries. Second, new departments were created during this period (e.g., Neuroscience, Genetics, or Biomedical Engineering), while others were phased out or dramatically shrunk (e.g., Anatomy). Third, the merging or survival of many departments are often a reflection of internal political struggles, rather than characteristics of the research conducted within them. For example, in some medical schools, orthopedic surgeons are in a separate department while in others, they are part of a large surgery department; In the basic sciences, many faculty would feel equally at home in cell biology, molecular biology, or biochemistry. Finally, three large departments (internal medicine, pediatrics, and surgery) tend to account for a large proportion of medical school employment, but their size masks enormous heterogeneity (e.g., neurosurgeons vs. cardio-thoracic surgeons; endocrinologists vs. infectious diseases specialists).

On substantive grounds, the production of abstract knowledge depends on the extent to which ideas are shared among researchers, and an important mechanism for sharing knowledge is direct collaboration through coauthorship. While physical proximity is an important determinant of the social interactions through which coworkers can learn from one another, intellectual distance (the extent to which the coworker skills overlap or complement each other) and social distance (the degree of intimacy that prevails between individuals) can matter as well. As a result, we are loath to entirely subsume human capital externalities under location considerations.

3.3 Coauthor Matching

To identify coauthors, we have developed a custom software program, the Stars/Colleagues Generator, or S/CG.⁶ The source of the publication data is PubMed, an online resource from the National Library of Medicine that provides fast, free, and reliable access to the biomedical research literature. In a first step, the S/CG downloads from the internet the entire set of English-language articles for a superstar, provided they are not letters to the editor, comments, or other “atypical” articles. These publications are used to compute raw and quality-adjusted publication counts for each superstar.

The meaning of coauthorship. In the life sciences, the number of coauthors is higher than in the social sciences (6 to 8 authors is typical). A robust social norm systematically assigns last authorship to the principal investigator, first authorship to the junior author who was responsible for the actual conduct of the investigation, and apportions the remaining credit to authors in the middle of the authorship list, generally as a decreasing function of the distance from the extremities of the list. We are weary of conferring “accidental” coauthors — those that result when the superstar is in the middle of the authorship list — too much value, since they need not entail actual pooling of knowledge or skills. As a result, we restrain the universe of potential coauthors to the set of publications in which the superstar appears in first or last position on the authorship list. From this set of publications, the S/CG strips out the list of coauthors, eliminates duplicate names, matches each coauthor with the faculty roster, and stores the identifier of every coauthor for whom a match is found. In a final step, the software queries PubMed for each validated coauthor, and generates publication counts as well as coauthorship variables for each superstar/colleague dyad, in each year.

The S/CG does not generate a match for each coauthors. Some coauthors are postdocs, technicians, or graduate students who do not go on to faculty positions within our period of observation; Other coauthors have positions in foreign institutions; others still publish under

⁶The complete specifications are described in a technical working paper (Stellman et al., 2005), while the S/CG software itself can be used by other researchers under a GNU license. Note that the SCG takes the faculty roster as an input; we are not authorized to share this data with third-parties. However, it can be licensed (for a fee) from AAMC, provided a local IRB gives its approval and a confidentiality agreement protects the anonymity of individual faculty members.

names that differ from the faculty roster listing (for instance by being inconsistent with the use of middle initials, suffixes, or hyphens). We are more worried about generating spurious matches, however.

Name matching. An important limitation in the matching process is that PubMed does not record authors' full names, nor does it record their institutional affiliation; it only keeps track of authors by using a combination of last name, two initials, and a suffix (where the suffix and the second initial fields can be empty). There are no unique author identifiers, and no possibility to account for the different name variations that a given author uses throughout his/her career. As a result, the S/CG can generate more than one roster match for a given PubMed author name, and the quality of these matches will depend directly on the relative frequency of last names in the population. We take a number of steps to deal with name matching-related issues. First, for the superstars, we have painstakingly crafted individual queries that take into account subject matter and affiliation to return an accurate set of publications.⁷ Second, we use information about the temporal sequence of coauthorships and graduation years to eliminate *a priori* implausible matches. Third, we use the relative frequencies of name/initial combinations from the roster to generate regression weights. We provide more details on the intricacies of the matching process in the Data Appendix.

3.4 Descriptive Statistics

When applied to our sample of 4,764 superstars, the S/CG software identifies 70,321 coauthors, a full 34.76% of the labor market. This translates into 46 coauthors per superstar on average (the median is 38) — the distribution is displayed in Figure 4. It is appropriate to describe the data at two different level, that of the individual peer, and that of the superstar/coauthor dyad. The software generates 206,875 dyads, for a total of 4,874,121 dyad-year observations between 1975 and 2003.

⁷These queries form part of the information that the software takes as an input. They can be up to 244 characters long (see Stellman et al. (2005) for more details).

Peer characteristics. The demographic characteristics for the peers differ from that of their superstar colleagues. The proportion of female faculty is higher (18% vs. 11%, see Table 8); They are also younger (the modal peer graduated in 1983, 13 years later than the modal superstar), more likely to hold an MD degree (55.44% vs. 45%, Table 9), and more likely to be affiliated with a clinical department (74.95% vs. 58.48%, Table 10). 3,907 (5.56%) of the peers are superstars themselves, 26,804 (38.12%) are NIH grantees but not superstars, and the remaining 39,610 (56.33%) are neither NIH grantees nor superstars. On Table 11, Panel A, one can see that the peers are not much less prolific than their superstar coauthors in terms of publications (2.05 vs 2.64), though the difference is more marked with quality-weighted publications (5.90 vs. 12.07).⁸ Finally, Table 6 shows that peers often coauthor with more than one superstar, although the distribution of superstar coauthors for peers is much more skewed than the distribution of peer coauthors for superstars.

Dyad characteristics. Out of 206,875 dyads in the sample, 49,260 (23.81%) correspond to peers and superstars who are co-located at least one year during the observation period (the number is 33,521 (16.20%) for dyad members located between one and ten miles apart, and 11,118 (5.37%) for dyad members located between ten and fifty miles apart). Table 12 provides evidence of variation in the intensity of coauthorship across dyads. Almost half of the dyads have a single instance of coauthorship. This leaves open the possibility of experimenting with more stringent thresholds to define a coauthorship treatment effect (e.g., “at least three coauthored publications, in three different years”).

⁸Note that peers’ publication counts are generated by simple PubMed queries (e.g., "schwartz sm"[au] for both Stephen M. Schwartz and Suzanne M. Schwartz), whereas the PubMed queries for the superstars are much more elaborate (e.g., "schwartz sm"[au] AND (screening OR risk factor OR cancer OR hutchinson[ad] OR incidence) NOT (macrophage OR "smooth muscle" OR array OR endothel* OR arterial OR "vascular wall" OR angiotensin OR atheroscl* OR aort*) AND 1986:2005[dp]). This will tend to mechanically inflate publication counts for peers with common names. The average publication counts weighted by name frequency are 1.53 (unadjusted) and 4.55 (quality-adjusted). Table 7 displays the distribution of name frequencies for the peers, as well as typical examples of last name/initial combinations corresponding to each level.

3.5 Econometric Modeling

The econometric tests we present aim to analyze the determinants of changes in research output produced independently of the superstar. In the case of publications and patents, this poses no particular problem — one simply restricts the publication or patent count to those articles or patents in which the superstar does not appear on the authorship/inventorship list. For NIH grants, apportioning output in this way is more difficult, since the CGAF dataset only lists principal investigators (PIs) for each grant. To the extent that superstars often appear as co-PIs on grants for which their coauthors are PIs, using grant output might lead us to overestimate the extent of peer effects. NIH grant awards, on the other hand, are measured much more precisely than publications, since a common set of individual identifiers is used in the CGAF and the faculty roster.

Basic specification. Our estimating equation relates output (without superstar i) in year t by peer j to characteristics of i , j and the research institutions to which they are affiliated:

$$\begin{aligned} y_{-i,jt} = & \beta_0 + \beta_1 DISTANCE_{ijt} + \beta_2 COAUTH_{ijt} \\ & + \beta_3 DISTANCE_{ijt} \times COAUTH_{ijt} + \beta_4 X_{ijt} + \delta_t + \gamma_{ij} + \varepsilon_{ijt} \end{aligned} \quad (1)$$

where *DISTANCE* denotes the geographic distance between i and j , *COAUTH* measures the intensity of coauthorship between i and j , and X is a vector of individual and institution characteristics. The δ_t 's stand for a full set of calendar year dummy variables, and the γ_{ij} 's correspond to dyad fixed effects, consistent with our approach to analyze changes in y following a professional transition by superstar i .

In practice, we do not enter *DISTANCE* in the specification, but instead use a set of seven indicator variables corresponding to (1) colocation; (2) no colocation, but less than 10 miles separating i and j ; (3) more than 10 miles but less than 50 miles (which we assume is an upper bound on commuting distance within a metropolitan area); (4) more than 50 miles but less than 100 miles; (5) more than 100 miles but less than 250 miles; (6) more than 250 miles but less than 500 miles; (7) more than 500 miles but less than 1,000 miles; and (8) more than 1,000 miles but less than 2,000 miles between i and j . The residual

category corresponds to collaborations in which dyad members are located more than 50 miles apart. To measure the intensity of coauthorship, we use three different measures. The first is the annual flow of coauthored articles between i and j ; the second is a coauthorship “regime” indicator that switches on following the onset of collaboration; the third correspond to the cumulative stock of coauthored articles between i and j up to year t . Although the easiest to interpret, the regime formulation of the coauthorship effect might not be the most meaningful, since the bulk of the dyads in the data coauthor only once.

Since all dyad members eventually coauthor, it would be wrong to interpret β_1 as a measure of the effect that co-located superstars have on their non-coauthor peers. Rather, the coefficient captures the effect of proximity to the star *before* the onset of collaboration. In our panel dataset, β_1 and β_2 are separately identified within dyad because i (and j) can move independently of the existence of a coauthoring relationship between them. Under the assumption of exogenous mobility, the causal effect of superstar exposure is $\beta_1 + \beta_2 + \beta_3$. We expect the signs of β_1 and β_2 to be positive. Since we assume that j will reallocate effort away from collaborations with other peers towards collaboration with i once the collaboration begins, β_3 could be negative. Of course, we are also interested in testing hypotheses regarding the relative magnitudes of β_1 and β_2 , and to explore the most salient dimensions of heterogeneity in these treatment effects, such as calendar time, gender, career stage, or scientific proximity.

Control variables. The dyad fixed effects control for many individual characteristics one would expect to influence research output, such as gender, degree, and scientific field (although changes in department affiliations are quite frequent). Explicit incentives for research in academia do depend on the career stage; given the shallow slope of post-tenure salary increases, Levin and Stephan (1991) suggest that levels of investment in research should vary over the career life cycle. To flexibly account for life cycle effects, we include a set of age dummy variables, where age measures the number of years since a scientist earned his/her doctoral degree (MD or PhD).⁹

⁹As is well known, it is not possible to separately identify period effects and life cycle effects in the “within” dimension of a panel, because one cannot observe two individuals at the same point in time that

Overhead costs recovered by universities from grants awarded by federal agencies are important determinants of inputs essential to the production of biomedical research, such as graduate students, post-doctoral fellows, technicians, laboratory space, and capital equipment. Therefore, we include the log of NIH funding for the peer’s medical school as a covariate in all specifications.

Econometric considerations. The number of articles published, patents applied for, or NIH grants awarded are examples of count dependent variables — non-negative integers with many zeros and ones. For example, 27.98% of the dyad-year observations in the data correspond to years of no publication output; the figure climbs to 89.21% for successful grant applications. Following a long-standing tradition in the study of scientific and technical change, we present conditional maximum likelihood estimates of eqn. [1] based on the fixed-effect poisson and negative binomial models developed by Hausman et al. (1984).

Productivity vs. output. We do not observe changes in the quantity of effort devoted to each of peer j ’s collaborations, only the output resulting from the allocation of effort across collaborations. Therefore, we cannot distinguish between the story in which j becomes genuinely more productive in her interactions with other peers following “exposure” to superstar i , from the story in which there are no productivity spillovers, but i shifts j ’s aspiration level in such a way that hours worked increase enough to generate higher output. Although the latter might not constitute an externality (since j presumably bears the cost of increased effort), it would nonetheless correspond to a genuine superstar effect.

Endogenous mobility. These estimates could be biased because of endogenous mobility, i.e., stars switching affiliations in anticipation of (or in reaction to) changes in the “peer environment,” for example because they expect a deterioration in the quality of their local peers, or observe an upward trend in the productivity of colleagues at an institution they eventually join. Conversely, within-dyad estimates could be downward biased if superstars

have the same (career) age but earned their degrees in different years. In practice, our specifications identify time and age effects through an ad hoc normalization: we omit *one* of the calendar year dummies and *one* of the age dummies (the models do not include a constant term). We do not report the corresponding estimates because they are sensitive to the choice of omitted category (Hall et al., 2005).

who provide the most spillovers to colleagues are also the most socially entrenched, and therefore face the highest costs of mobility. To some degree, the extent of bias can be gauged by examining the determinants of professional transitions among superstars. For example, we do observe the number of local coauthors for each superstar, as well as recent trends in the funding at both the origin and destination institutions. To provide a causal interpretation of these estimates, genuine exogenous variation in the propensity to move is required, however. This is on our agenda. Potential instruments include the presence of a faculty spouse for the superstar (as well as a number of household characteristics of the couple, such as the extent of coauthorship within the couple, and the difference in talent between the superstar and his/her spouse). We could also use an indicator for the presence of children between the age of 10 and 18 in the superstar’s household — parents might be reluctant of destroying the social networks of their teenage children.

4 Results

Below, we restrict our attention to a select group of superstars: 976 individuals whose cumulative grant amount between 1977 and 2003 lie above the 99th percentile of the distribution of funding (both including, and excluding center grants). We impose this restriction in order to speed up the estimation of the models.

We begin by presenting results for specifications in which the location effect has been broken down into 7 categories (Table 13). All estimates are presented in the form of incidence rate ratios; the formula $(e^\beta - 1) \times 100\%$ (where β denotes an estimated coefficient) provides a number directly interpretable in terms of elasticity. Model (1), for instance, implies that becoming co-located with a star is associated with a 13% increase in publication output, whereas becoming located less than 10 miles away only increases output by 7.6%, and becoming located between 10 and 50 miles away increases output by 8.8%. Further away, the impact of superstar is much smaller in magnitude, and even negative in the 50 to 250 miles range. Model (2) focuses on the coauthorship treatment effect alone, which is in the order of 5%. Very similar results are obtained in Model (3), which combines the location

and coauthorship effects. Model (4) adds a full set of interaction terms between location and coauthorship. In this model, the location effects measure the impact of superstar exposure *before* the onset of a scientific collaboration with the star. The patterns for the pure location effect are identical to those in Model (3), but the magnitudes are higher — 25.9% higher publication output for co-located peers, for instance. The coauthorship effect is also higher, on the order of 9%. However, this specification also provides evidence of partial substitution between collaborations involving the star and collaborations with other researchers. The interaction effect between co-location and coauthorship is both negative and highly statistically significant. The total effect of superstar exposure can be obtained by adding each location effect with its corresponding interaction and the pure coauthorship effect. These calculations yield 17.4% for co-location, 11.6% for the “less than 10 miles” range, and 13.6% for the “10 to 50 miles” range. Beyond 50 miles, the total effect hovers slightly above or below 10%: the location and substitution effects precisely cancel one another.

One problem for the dyad-level analysis is that a given peer will often coauthor with more than one superstar. The estimated effects might be correlated across these multiple relationships. Model (5) investigates this possibility by adding to the specification a count of other ongoing collaborations with superstars for the peer. This has an important effect on the results. First, the magnitude of the location effects is dampened. Second, the coauthorship effect disappears, indicating that it is difficult to separately identify the pure coauthorship effects corresponding to multiple collaborations.

Table 14 performs a similar exercise, but defines the coauthorship effect in two other ways. In columns (1) and (2), coauthorship is measured as the yearly flow of coauthored articles; in columns (3) and (4), coauthorship is measured as the cumulative stock of coauthored articles up to the current year. The results are broadly in line with those of Table 13, with the notable exception that the stock formulation of the coauthorship effect enables us to separately identify the effect of coauthorship with the focal superstar from the effect of coauthorship with other superstars.

In Table 15, we focus on a different measure of output: the number of NIH grants awarded in a year. We focus on the number of grants rather than dollar amounts because grants are typically awarded for a period of 5 years, and disbursed in equal yearly amounts over this period; Only the first of these payments is indicative of successful grantsmanship, however.

Because a large number of our peers never receive a grant from the NIH, the sample size is smaller in these analyses. As in the case of publications, we find evidence of an effect of superstars, but no clear pattern really emerges from the data regarding the effect of location, except maybe in the 50 to 100 miles range. In contrast with the results of Tables 13 and 14, the substitution effects appear to increase with distance! At this stage, we do not really understand what might explain these results, but we remind the reader that grants are a problematic measure of output to the extent that we do not observe collaboration patterns on grants.

Table 16 focuses on the evolution of the returns to co-location over the period 1975-2003, by interacting the co-location effect with a series of binary variables corresponding to 5-year intervals. Model (1), which only includes co-location effects, indicates that the spillover benefits of co-location remained stable until the late 1980s, but declined precipitously afterwards. In fact, towards the end of the 1990s, the co-location effect disappears, and even turns negative in the period 2001-2003. In Model (2), we see that the spillover benefits of coauthorship follow a similar pattern, except that it is even more pronounced. Model (3) shows that entering the pure co-location and pure coauthorship effects simultaneously does not modify these conclusions. These results are puzzling. One would expect the diffusion of the internet to diminish the role of geographic distance as a factor influencing the extent of spillovers, but not to the extent of making physical proximity irrelevant, or a hindrance. As for the benefits from coauthorship, there is no obvious interpretation for the negative effects implied by our estimates after 1995. Remember, however, that it is difficult to estimate the pure coauthorship effects for a specific dyad (Table 13 and 14).

Model (4) adds to the specification three-way interaction terms between co-location, coauthorship, and time. The results imply that, in the 1990s, superstars provide much less

benefits to their co-located peers before they become coauthors, but that they benefits they do provide after the beginning of their collaboration are enough to overcome any substitution effect. Toward the end of the period, the sum of the pure coauthorship, pure co-location, and interaction effects is close to 0, but the interaction effect by itself is strongly positive and significant. The results are consistent with a model in which superstars became more selective over time in their choice of local coauthors, but also have a stronger connection with those they associate with, at least after the collaboration has begun to produce research output.

5 Conclusion

In this paper, we have presented an analysis of scientific collaborations in the academic life sciences, focusing more particularly on the impact of superstars on their peers. Using the movements of these superstars across institutions, we uncover evidence consistent with fairly large spillover effects. These benefits of exposure to star talent can be decomposed into location effects and collaboration effects. We can identify these effects separately because (1) not all collaborations involve co-located colleagues, and (2) collaborators might be co-located before their relationship yields a visible audit trail in the form of co-authored publications. We also provide evidence of substitution towards output jointly produced with the superstar, but this substitution only cancels about half of the sum of the “pure location” and “pure coauthorship” effects.

In related analyses, we find that the spillover effects associated with location have declined markedly over time. This is consistent with the idea that improved information and communication technologies, such as the rise of the internet, have reduced the cost of distant collaborations.

At this stage, our conclusions must remain tempered for two reasons. First, we have not dealt at all with the issue of endogenous mobility. As stated earlier, we can pursue a variety of avenues to explore the extent to which this biases our results. Second, even if we assume mobility to be exogenous, it is unclear whether our estimates reflect genuine externalities. We observe output, not productivity; an interpretation of our results is that superstars shift their

peer's aspiration level, thus triggering more effort, even in collaborations that do not involve the star. While this corresponds to a real effect of the star, the term externality should not apply because the peer presumably bears the cost of increased effort. We must also resist calling these effects externalities because we document the benefits of collaborations, but scientific coauthorships also entail coordination costs. These could be borne by the peer in the form of lower wages, or by the star, who might divert some of his/her effort towards mentorship activities. Yet, we suspect that these spillovers are not fully internalized by the scientific labor market. In academia, pay exhibits much less dispersion than talent, and labor market frictions limit superstars' ability to extract their full marginal product.

The data presented here provide a platform to explore the “technology” of knowledge spillovers. We could extend the analysis to include other measures of distance, such as “scientific distance” and “social distance.” Do collaborations involving complementary skill sets yield higher output than those pooling the effort of researchers with similar knowledge? Do further “degrees of coauthor separation” also yield spillover benefits? These are only examples of a number of fascinating questions that these data should enable us to answer.

References

- Acemoglu, Daron. "A Microfoundation for Social Increasing Returns in Human Capital Accumulation." *Quarterly Journal of Economics* 111, no. 3 (1996): 779-804.
- Acemoglu, Daron, and Joshua Angrist. "How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws." In *NBER Macroeconomics Annual 2000*, edited by Ben S. Bernanke and Kenneth S. Rogoff, 9-59. Cambridge, MA: The MIT Press, 2001.
- Acemoglu, Daron, and Jörn-Steffen Pischke. "The Structure of Wages and Investment in General Training." *Journal of Political Economy* 107, no. 3 (1999): 539-72.
- . "Why Do Firms Train? Theory and Evidence." *Quarterly Journal of Economics* 113, no. 1 (1998): 79-119.
- Adams, James D., and Adam B. Jaffe. "Bounding the Effects of R&D: An Investigation Using Matched Establishment-Firm Data." *RAND Journal of Economics* 27, no. 4 (1996): 700-21.
- Allison, Paul D., and J. Scott Long. "Departmental Effects on Scientific Productivity." *American Sociological Review* 55, no. 4 (1990): 469-78.
- . "Interuniversity Mobility of Academic Scientists." *American Sociological Review* 52, no. 5 (1987): 643-52.
- Azoulay, Pierre, Waverly Ding, and Toby Stuart. "The Determinants of Faculty Patenting Behavior: Demographics or Opportunities?" NBER Working Paper #11348 (2005).
- Cech, Thomas R. "Fostering Innovation and Discovery in Biomedical Research." *Journal of the American Medical Association* 294, no. 11 (2005): 1390-93.
- Cockburn, Iain M., and Rebecca M. Henderson. "Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery." *Journal of Industrial Economics* 46, no. 2 (1998): 157-82.
- Conley, Timothy, and Christopher Udry. "Social Learning through Networks: The Adoption of New Agricultural Technologies in Ghana." *American Journal of Agricultural Economics* 83, no. 3 (2001): 668-73.
- Costa, Dora L., and Matthew E. Kahn. "Cowards and Heroes: Group Loyalty in the American Civil War." *Quarterly Journal of Economics* 118, no. 2 (2003): 519-48.
- . "Power Couples: Changes in the Locational Choice of the College Educated, 1940-1990." *Quarterly Journal of Economics* 115, no. 4 (2000): 1287-315.
- de Solla Price, Derek J. *Little Science, Big Science*. New York: Columbia University Press, 1986.

- Fafchamps, Marcel, Sanjeev Goyal, and Marco van de Leij. "Scientific Networks and Co-Authorship." Working paper, University of Oxford (2004).
- Hall, Bronwyn H., Jacques Mairesse, and Laure Turner. "Identifying Age, Cohort and Period Effects in Scientific Research Productivity: Discussion and Illustration Using Simulated and Actual Data on French Physicists." NBER Working Paper #11739 (2005).
- Henderson, Rebecca, Luigi Orsenigo, and Gary P. Pisano. "The Pharmaceutical Industry and the Revolution in Molecular Biology: Interactions among Scientific, Institutional, and Organizational Change." In *Sources of Industrial Leadership*, edited by David C. Mowery and Richard R. Nelson, 267-311. New York: Cambridge University Press, 1999.
- Hoxby, Caroline. "Peer Effects in the Classroom: Learning from Gender and Race Variation." NBER Working Paper #7867 (2000).
- Ichino, Andrea, and Giovanni Maggi. "Work Environment and Individual Background: Explaining Regional Shirking Differentials in a Large Italian Firm." *Quarterly Journal of Economics* 115, no. 3 (2000): 1057-90.
- Jaffe, Adam B. "Technological Opportunity and Spillovers from R&D: Evidence from Firms' Patents, Profits, and Market Value." *American Economic Review* 76, no. 5 (1986): 984-1001.
- Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* 108, no. 3 (1993): 577-98.
- Jones, Benjamin F. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" NBER Working Paper #11360 (2005).
- Kim, E. Han, Adair Morse, and Luigi Zingales. "Are Elite Universities Losing Their Competitive Edge?" Working Paper, University of Michigan (2005).
- Lacetera, Nicola, Iain M. Cockburn, and Rebecca Henderson. "Do Firms Change Their Capabilities by Hiring New People? A Study of the Adoption of Science-Driven Drug Research." In *Advances in Strategic Management*, edited by Joel Baum and Anita McGahan, 133-59. New York: Elsevier, 2004.
- Lazear, Edward P. "Raids and Offer Matching." *Research in Labor Economics* 8, no. A (1986): 141-65.
- Levin, Sharon G., and Paula E. Stephan. "Research Productivity over the Life Cycle: Evidence for Academic Scientists." *American Economic Review* 81, no. 1 (1991): 114-32.
- Lotka, Alfred J. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* 16 (1926): 317-23.

- Lucas, Robert E., Jr. "On the Mechanics of Economic Development." *Journal of Monetary Economics* 22, no. 1 (1988): 3-42.
- Mairesse, Jacques, and Laure Turner. "Measurement and Explanation of the Intensity of Co-Publication in Scientific Research: An Analysis at the Laboratory Level." NBER Working Paper #11172 (2005).
- Manski, Charles F. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60, no. 3 (1993): 531-42.
- Marshall, Alfred. *Principles of Economics*. New York: MacMillan, 1890.
- Merton, Robert. "The Normative Structure of Science." In *The Sociology of Science: Theoretical and Empirical Investigations*, edited by Robert Merton. Chicago: University of Chicago Press, 1973.
- Metzger, Norman, and Richard N. Zare. "Interdisciplinary Research: From Belief to Reality." *Science* 283, no. 5402 (1999): 642-43.
- Moretti, Enrico. "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions." *American Economic Review* 94, no. 3 (2004): 656-90.
- Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. "The Allocation of Talent: Implications for Growth." *Quarterly Journal of Economics* 106, no. 2 (1991): 503-30.
- Nelson, Richard R., and Edmond S. Phelps. "Investment in Humans, Technological Diffusion and Economic Growth." Cowles Foundation Discussion Paper #189, Yale University (1965).
- Rauch, James E. "Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities." *Journal of Urban Economics* 34, no. 3 (1993): 380-400.
- Romer, Paul M. "Endogenous Technological Change." *Journal of Political Economy* 98, no. 5 (1990): S71-S102.
- Rosenblat, Tanya S., and Markus M. Mobius. "Getting Closer or Drifting Apart?" *Quarterly Journal of Economics* 119, no. 3 (2004): 971-1009.
- Sacerdote, Bruce. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116, no. 2 (2001): 681-704.
- Schultz, Theodore. *The Economic Value of Education*. New York: Columbia University Press, 1967.
- Sobecks, Nancy, Amy C. Justice, Susan Hinze, Heidi Taylor Chirayath, Rebecca J. Lasek, Mary-Margaret Chren, John Aucott, Barbara Juknialis, Richard Fortinsky, Stuart Youngner, and C. Seth Landefeld. "When Doctors Marry Doctors: A Survey Exploring the Professional and Family Lives of Young Physicians." *Annals of Internal Medicine* 130, no. 4 (1999): 312-19.

Stellman, Andrew, Pierre Azoulay, and Joshua Graff Zivin. "Stars/Colleagues Generator: Software Specifications." Working Paper, Columbia University (2005).

Weitzman, Martin L. "Recombinant Growth." *Quarterly Journal of Economics* 113, no. 2 (1998): 331-60.

Zucker, Lynne G., Michael R. Darby, and Marilyn B. Brewer. "Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises." *American Economic Review* 88, no. 1 (1998): 290-306.

Table 1
Distribution of Superstar by Achievement Criteria

	NIH Funding Superstar	NIH Funding Superstar (excl. center grants)	Highly Cited	HHMI	Patent Superstar
NIH funding Superstar (>\$14,075,000)	3,123	1,990	421	105	74
NIH funding, excluding center grants (>\$10,468,000)	1,990	2,948	400	98	79
Highly Cited	421	400	776	104	66
HHMI	105	98	104	540	25
Patent Superstar (#>17)	74	79	66	25	127

Table 2
Descriptive Statistics, Superstar Sample

	N	Mean	Std. Dev.	Min	Max
Nb. of positions	4,764	1.881	1.000	1	8
Female	4,764	0.118	0.323	0	1
Nobel Laureate	4,764	0.008	0.090	0	1
Lasker Award Winner	4,764	0.014	0.117	0	1
Inst. of Medicine	4,764	0.107	0.310	0	1
Nat. Acad. of Sciences	4,764	0.090	0.287	0	1
NIH Career Funding	4,764	\$25,498,738	\$25,598,272	0	\$462,601,248
NIH Career Funding, excl. center grants	4,764	\$14,663,186	\$14,122,324	0	\$330,966,944
Pub. Count (first or last author)	4,764	65.907	49.667	0	450
Wghtd Pub. Count (first or last author)	4,764	301.090	287.245	0	2883.665
Career Patents	4,764	2.097	6.252	0	202
NIH Grants, annual flow	142,542	0.413	0.696	0	12
NIH Grants, annual flow, excl. center grants	142,542	0.367	0.666	0	12
NIH Funding, annual flow	142,542	\$851,128	\$1,502,743	0	\$84,952,536
NIH Funding, annual flow, excl. center grants	142,542	\$489,368	\$870,649	0	\$38,857,284
Pub. count, annual flow (first or last author)	118,839	2.641	2.824	0	40
Wghtd Pub. count, annual flow (first or last author)	118,839	12.066	16.366	0	285.634
Nb. of Coauthors, annual flow	137,078	3.245	4.341	0	64
Nb. of Coauthors, colocated, annual flow	137,078	0.879	1.748	0	28
Nb. of Coauthors, 0 < distance < 10 miles, annual flow	137,078	0.477	1.492	0	42
Nb. of Coauthors, 10 < distance < 50 miles, annual flow	137,078	0.108	0.555	0	21
Nb. of Coauthors, 50 miles < distance, annual flow	137,078	1.729	2.714	0	52

Table 3
Top 20 Research Institutions where Superstars Reside

Rank	School	Nb. Scientist/Year Obs.	Prop.
1	Harvard Medical School	12,302	8.63%
2	Johns Hopkins University School of Medicine	5,218	3.66%
3	University of Washington School of Medicine	5,040	3.54%
4	UCSF School of Medicine	4,924	3.45%
5	Yale University School of Medicine	4,491	3.15%
6	Columbia University College of Physicians & Surgeons	4,410	3.09%
7	University of Pennsylvania School of Medicine	4,335	3.04%
8	Stanford University School of Medicine	3,891	2.73%
9	Washington University in Saint Louis School of Medicine	3,785	2.66%
10	UCLA School of Medicine	3,509	2.46%
11	Joan & Sanford I. Weill Medical College — Cornell University	2,907	2.04%
12	Duke University School of Medicine	2,843	1.99%
13	Albert Einstein College of Medicine — Yeshiva University	2,728	1.91%
14	UCSD School of Medicine	2,589	1.82%
15	University of Michigan Medical School	2,460	1.73%
16	University of Chicago Pritzker School of Medicine	2,379	1.67%
17	NIH	2,284	1.60%
18	University of Minnesota Medical School	2,280	1.60%
19	New York University School of Medicine	2,193	1.54%
20	Baylor College of Medicine	2,179	1.53%
Total Top 20		76,747	53.84%
Total		142,542	100.00%

Table 4
Main Department Breakdown, Superstars of Medicine

	Freq.	Prop. of Basic Scientists	Prop. of Total
Biochemistry	491	24.82%	10.31%
Microbiology	281	14.21%	5.90%
Pathology	272	13.75%	5.71%
Pharmacology	223	11.27%	4.68%
Physiology	195	9.86%	4.09%
Genetics	155	7.84%	3.25%
Neuroscience	146	7.38%	3.06%
Cell Biology	142	7.18%	2.98%
Anatomy	50	2.53%	1.05%
Biomedical Engineering	17	0.86%	0.36%
Other Basic Sciences	6	0.30%	0.13%
Total, Basic Science Departments	1,978		41.52%
	Freq.	Prop. of Clinical Scientists	Prop. of Total
Internal Medicine	1,225	43.97%	25.71%
Psychiatry	360	12.92%	7.56%
Pediatrics	297	10.66%	6.23%
Public Health & Preventive Medicine	235	8.44%	4.93%
Neurology	169	6.07%	3.55%
Surgery	154	5.53%	3.23%
Radiology	116	4.16%	2.43%
Ophthalmology	78	2.80%	1.64%
Obstetrics/Gynecology	51	1.83%	1.07%
Otolaryngology	31	1.11%	0.65%
Dermatology	30	1.08%	0.63%
Anesthesiology	18	0.65%	0.38%
Community & Family Medicine	17	0.61%	0.36%
Dentistry	3	0.11%	0.06%
Physical Medicine & Rehabilitation	2	0.07%	0.04%
Total, Clinical Departments	2,786		58.48%
Grand Total	4,764		100.00%

Table 5
Job Mobility, Superstars of Medicine

	Freq.	Proportion
No moves	2,128	44.67%
One transition	1,524	31.99%
Two Transitions	767	16.10%
Three Transitions	262	5.50%
Four or more transitions	83	1.74%

Table 6
Number of Superstar Coauthors per Peer

	Freq.	Proportion
1	30,705	43.66%
2	14,970	21.29%
3	8,202	11.66%
4	5,025	7.15%
5	3,132	4.45%
6	2,101	2.99%
7	1,431	2.03%
8	1,018	1.45%
9	798	1.13%
10	551	0.78%
11	431	0.61%
12	398	0.57%
13	263	0.37%
14	213	0.3%
15	157	0.22%
16	128	0.18%
17 or more (99 th Percentile)	798	1.13%
Total	70,321	100%

Table 7
Distribution of Name Frequencies (last name/2 initials)

	Freq.	Proportion	Example
1	54,317	77.24%	Lechleiter JD
2	8,663	12.32%	Weinstein SL
3	3,493	4.97%	Scott WJ
4	1,597	2.27%	Patel SC
5	721	1.03%	Young RC
6	413	0.59%	Brown RD
7	359	0.51%	Anderson RJ
8	160	0.23%	Cohen SM
9	124	0.18%	Jones JE
10	99	0.14%	Miller MJ
11	100	0.14%	Kim HS
12	41	0.06%	Gupta S
13	33	0.05%	Smith DM
14	33	0.05%	Wu Y
15	31	0.04%	Johnson JA
16	34	0.05%	Chen J
17	22	0.03%	Kumar A
19	13	0.02%	Smith JA
20	16	0.02%	Li Y
24	24	0.03%	Zhang Y
36	28	0.04%	Wang Y
Total	70,321	100%	

Table 8
Peer Gender Distribution

	Freq.	Proportion
Male	53,070	75.47%
Ambiguous	4,504	6.4%
Female	12,747	18.13%
Total	70,321	100%

Gender is assigned probabilistically based on first names, using the genderizer software of Language Analysis Systems, Inc. The “male” (resp. “female”) category corresponds to first names for which the actual proportion of males (resp. females) bearing this name is greater than 80%.

Table 9
Peer Degree Distribution

	Freq.	Proportion
MD	38,986	55.44%
PhD	24,773	35.23%
MD/PhD	6,062	8.62%
Other Health Doctorate (DDS, DVM, PharmD, DrPH, etc.)	500	0.71%
	70,321	100.00%

Table 10
Main Department Affiliation, Peers

		Freq.	Prop. of Basic/Clin. Scientists	Prop. of Total
Basic Science Departments	Anatomy	1,667	9.64%	2.37%
	Biochemistry	3,016	17.45%	4.29%
	Microbiology	2,049	11.85%	2.91%
	Pathology	4,829	27.93%	6.87%
	Pharmacology	2,075	12.00%	2.95%
	Physiology	2,069	11.97%	2.94%
	Other Basic Sciences	1,582	9.15%	2.25%
	Total	17,287	100.00%	24.58%
Clinical Departments	Anesthesiology	1,938	3.68%	2.76%
	Dermatology	529	1.00%	0.75%
	Family Medicine	900	1.71%	1.28%
	Internal Medicine	18,531	35.16%	26.35%
	Neurology	2,557	4.85%	3.64%
	Obstetrics & Gynecology	2,185	4.15%	3.11%
	Ophthalmology	1,476	2.80%	2.10%
	Orthopaedics	762	1.45%	1.08%
	Otolaryngology	682	1.29%	0.97%
	Pediatrics	6,397	12.14%	9.10%
	Physical Medicine and Rehabilitation	343	0.65%	0.49%
	Psychiatry	4,939	9.37%	7.02%
	Public Health and Preventative Medicine	1,856	3.52%	2.64%
	Radiology	4,052	7.69%	5.76%
Surgery	5,557	10.54%	7.90%	
	Total	52,704	100.00%	74.95%
	Miscellaneous	330		0.47%
	Grand Total	70,321		100.00%

Table 11
Descriptive Statistics, Peers and Superstar-Peer Dyads

	Obs.	Mean	Std. Dev.	Min.	Max.
<i>Panel A — Peer-Year Level</i>					
Pub. Count, First or Last Author	1,508,789	2.049	5.533	0	222
Wghtd Pub. Count, First or Last Author	1,508,789	5.895	16.982	0	705.081
Number of NIH Grants	1,508,789	0.090	0.339	0	12
Number of NIH Grants, excl. Center Grants	1,508,789	0.086	0.329	0	12
NIH Funding, annual flow	1,508,789	\$113,474	\$479,849	0	\$59,783,728
NIH Funding, annual flow, excl. Center grants	1,508,789	\$82,543	\$313,096	0	\$38,857,284
<i>Panel B — Dyad-Year Level</i>					
Pub. Count, First or Last Author, No Superstar	4,874,121	4.291	11.901	0	222
Wghtd Pub. Count, First or Last Author, No Superstar	4,874,121	13.040	36.630	0	705.081
Number of Coauthorships, Annual Flow	4,874,121	0.164	0.666	0	30
Coauthorship Regime	4,874,121	0.510	0.500	0	1
Number of Coauthorships, Cumulative Stock	4,874,121	1.995	5.084	0	298
Star & Peer, Colocated	4,874,121	0.151	0.358	0	1
Star & Peer, less than 10 miles apart	4,874,121	0.097	0.295	0	1
Star & Peer, 10<distance<50 miles	4,874,121	0.028	0.165	0	1
Star & Peer, distance greater than 50 miles	4,874,121	0.637	0.481	0	1

Table 12
Number of Total Coauthorships per Superstar-Peer Dyad

	Freq.	Prop.	Prcntl.
One Coauthorship	90,622	43.81%	43.81%
Two Coauthorships	41,488	20.05%	63.86%
Three Coauthorships	17,051	8.24%	72.10%
Between 4 and 8 Coauthorships	36,340	17.57%	89.67%
Between 9 and 13 Coauthorships	10,714	5.18%	94.85%
Between 14 and 31 Coauthorships	8,490	4.10%	98.95%
Between 32 and 298 Coauthorships	2,170	1.05%	100.00%
Total	206,875	100.00%	

Figure 1
Career NIH Funding for the Superstars of Medicine

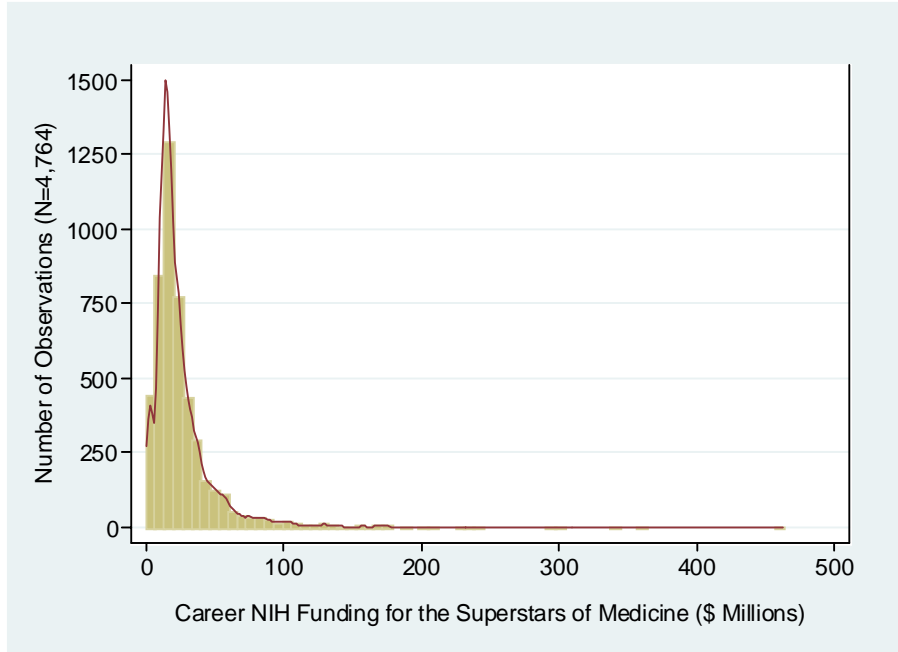


Figure 2
Career Publication Count for the Superstars of Medicine

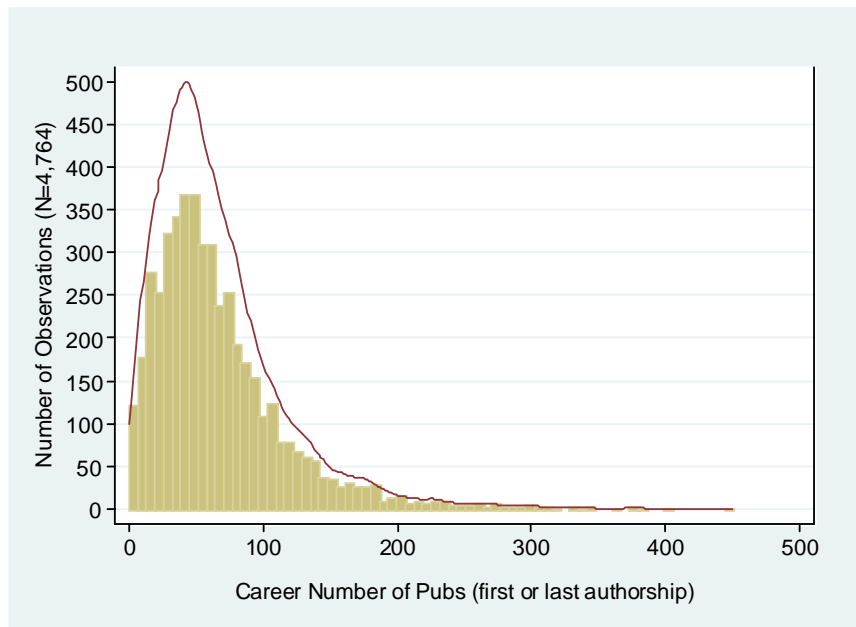


Figure 3
Career Number of Patents for the Superstars of Medicine

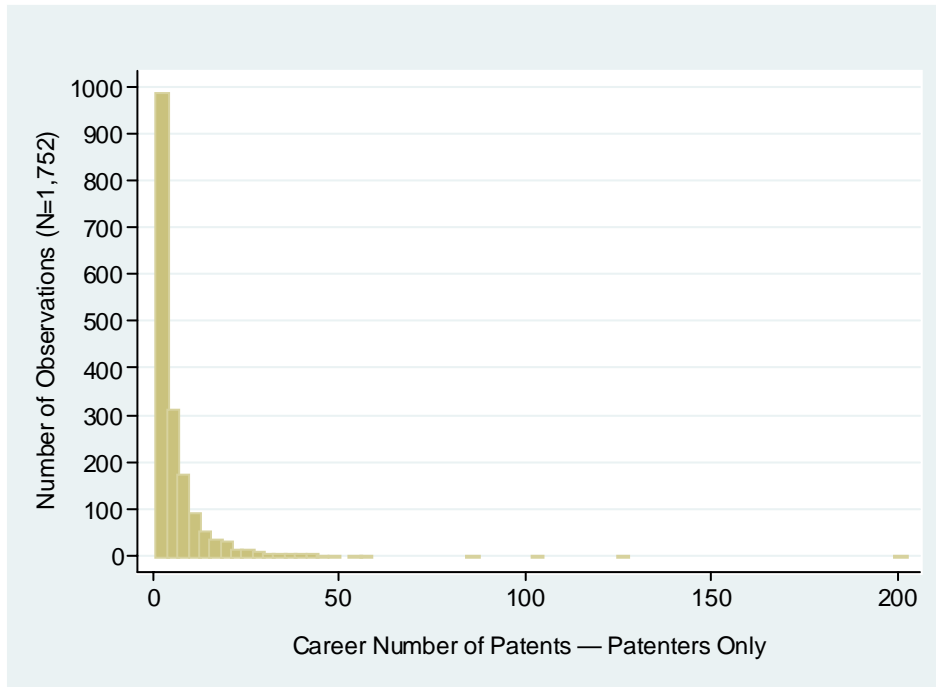


Figure 4
Number of Faculty Coauthors per Superstar

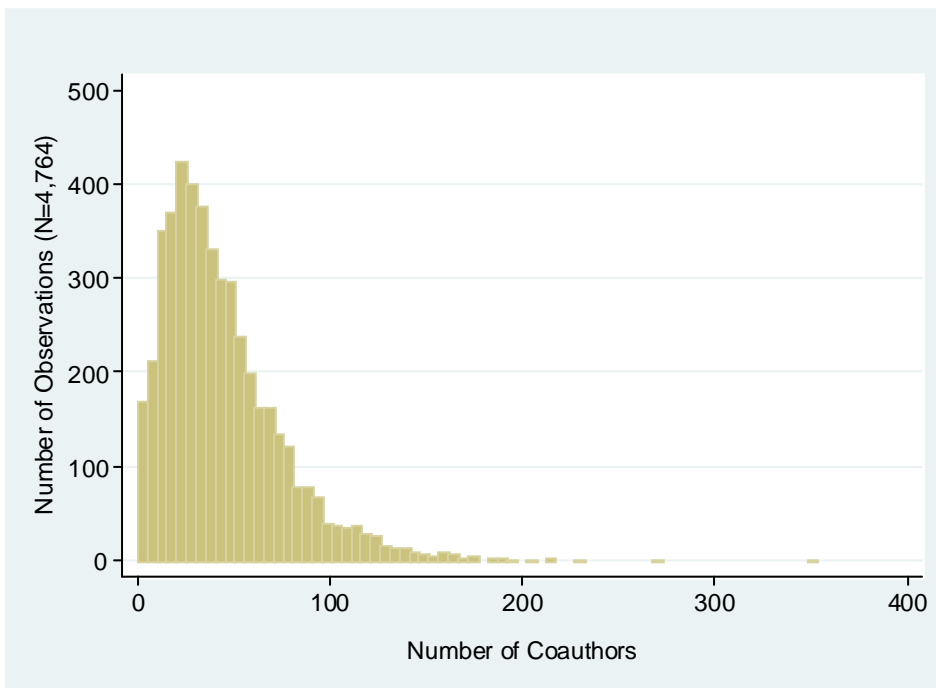


Table 13

Effect of superstar exposure on peer publication output. Conditional MLE estimates for the fixed effect Negative Binomial Model. The dyads considered are those corresponding to superstars above the the 99th prcntl. of the Cumulative NIH grant distribution. The estimates reported correspond to incidence rate ratios.

	(1)	(2)	(3)	(4)	(5)
Co-located	1.129** [19.327]		1.125** [18.693]	1.259** [30.252]	1.171** [20.865]
Less than 10 miles apart	1.076** [9.913]		1.073** [9.606]	1.155** [16.656]	1.090** [10.008]
Between 10 & 50 miles	1.088** [7.143]		1.086** [6.991]	1.162** [10.497]	1.092** [6.183]
Between 50 & 100 miles	0.961** [3.589]		0.960** [3.618]	0.927** [5.692]	0.945** [4.321]
Between 100 & 250 miles	0.988 [†] [1.718]		0.988 [†] [1.673]	0.965** [4.213]	0.966** [4.147]
Between 250 & 500 miles	1.027** [4.117]		1.027** [4.055]	1.031** [3.952]	1.000 [0.032]
Between 500 & 1000 miles	1.020** [3.563]		1.021** [3.572]	1.005 [0.751]	0.983** [2.629]
Between 1000 & 2000 miles	1.043** [7.224]		1.043** [7.219]	1.042** [5.955]	1.020** [2.926]
Coauthorship, Regime		1.047** [14.991]	1.043** [13.789]	1.089** [13.132]	0.998 [0.388]
Nb. Other Star Coauthors					1.025** [181.524]
Co-located × Coauthorship Reg.				0.826** [23.203]	0.903** [12.603]
Less than 10 miles × Coauthorship Reg.				0.872** [14.939]	0.940** [6.966]
Between 10 & 50 miles × Coauthorship Reg.				0.885** [7.925]	0.959** [2.781]
Between 50 & 100 miles × Coauthorship Reg.				1.078** [5.049]	1.076** [5.211]
Between 100 & 250 miles × Coauthorship Reg.				1.049** [5.009]	1.064** [6.773]
Between 250 & 500 miles × Coauthorship Reg.				0.989 [1.274]	1.026** [2.996]
Between 500 & 1000 miles × Coauthorship Reg.				1.030** [3.817]	1.059** [7.857]
Between 1000 & 2000 miles × Coauthorship Reg.				0.998 [0.214]	1.024** [3.096]
Ln(Med. School NIH Funding)	1.031** [17.004]	1.035** [19.701]	1.031** [16.884]	1.031** [17.034]	1.029** [16.383]
Observations	960,091	960,091	960,091	960,091	960,091
Number of dyads	44,913	44,913	44,913	44,913	44,913
Log Likelihood	-1,399,568	-1,399,749	-1,399,473	-1,398,775	-1,387,280
Sum of Spillover Coeffs. = 0			541.682**	267.112**	50.561**

Notes:

- (1) Absolute value of z statistics in brackets.
(2) Dependent variable is number of publications w/o the superstar, peer first or last author.
(2) All models include a set of calendar year dummies and career age dummies (coefficients not reported).
(3) [†]significant at 10%; ^{*}significant at 5%; ^{**}significant at 1%.

Table 14

Effect of superstar exposure on peer publication output. Conditional MLE estimates for the fixed effect Negative Binomial Model. The dyads considered are those corresponding to superstars above the the 99th prcntl. of the Cumulative NIH grant distribution. The estimates reported correspond to incidence rate ratios.

	<i>Coauthorship Flow</i>		<i>Coauthorship Stock</i>	
	(1)	(2)	(3)	(4)
Co-located	1.141** [20.77]	1.125** [18.83]	1.161** [22.493]	1.131** [18.72]
Less than 10 miles apart	1.087** [11.27]	1.073** [9.62]	1.096** [11.930]	1.072** [9.07]
Between 10 & 50 miles	1.095** [7.63]	1.081** [6.63]	1.101** [7.732]	1.075** [5.88]
Between 50 & 100 miles	0.961** [3.59]	0.967** [3.09]	0.950** [4.326]	0.963** [3.20]
Between 100 & 250 miles	0.987† [1.77]	0.990 [1.49]	0.984† [2.120]	0.987† [1.81]
Between 250 & 500 miles	1.027** [4.08]	1.011† [1.77]	1.029** [4.127]	1.015* [2.13]
Between 500 & 1000 miles	1.020** [3.43]	1.007 [1.27]	1.010† [1.690]	1.001 [0.21]
Between 1000 & 2000 miles	1.042** [7.08]	1.032** [5.42]	1.044** [7.038]	1.041** [6.62]
Coauthorship, Flow or Stock	1.027** [5.20]	0.998 [0.44]	1.011** [14.094]	1.005** [6.39]
Nb. of Other Star Coauthors, Flow or Stock		1.054** [245.29]		1.005** [132.19]
Co-located ×	0.951** [8.55]	0.959** [7.39]	0.993** [8.701]	0.994** [7.51]
Coauthorship Flow/Stock				
Less than 10 miles ×	0.942** [9.24]	0.957** [7.11]	0.995** [5.928]	0.997** [3.80]
Coauthorship Flow/Stock				
Between 10 & 50 miles ×	0.947** [5.04]	0.970* [2.97]	0.996* [2.392]	1.000 [0.11]
Coauthorship Flow/Stock				
Between 50 & 100 miles ×	0.988 [1.04]	1.001** [0.13]	1.007** [3.497]	1.007** [3.36]
Coauthorship Flow/Stock				
Between 100 & 250 miles ×	0.992 [1.06]	1.000* [0.06]	1.003* [2.283]	1.004** [2.93]
Coauthorship Flow/Stock				
Between 250 & 500 miles ×	0.989 [1.53]	0.982 [2.63]	0.999 [0.718]	1.001 [0.82]
Coauthorship Flow/Stock				
Between 500 & 1000 miles ×	0.994 [0.98]	0.994** [0.92]	1.006** [6.127]	1.006** [6.27]
Coauthorship Flow/Stock				
Between 1000 & 2000 miles ×	0.996 [0.58]	1.004 [0.60]	0.999 [0.598]	0.995** [4.88]
Coauthorship Flow/Stock				
Ln(Med. School NIH Funding)	1.031** [17.03]	1.023** [12.73]	1.031** [17.148]	1.030** [16.53]
Observations	960,091	960,091	960,091	960,091
Number of dyads	44,913	44,913	44,913	44,913
Log Likelihood	-1,399,434	-1,377,643	-1,399,129	-1,392,360
Sum of Spillover Coeffs. = 0	281.08**	134.97**	539.680**	348.970**

Notes:

- (1) Absolute value of z statistics in brackets.
- (2) Dependent variable is number of publications w/o the superstar, peer first or last author.
- (2) All models include a set of calendar year dummies and career age dummies (coefficients not reported).
- (3) †significant at 10%; *significant at 5%; **significant at 1%.

Table 15

Effect of superstar exposure on peer grant output. Conditional MLE estimates for the fixed effect Negative Binomial Model. The dyads considered are those corresponding to superstars above the the 99th prcntl. of the Cumulative NIH grant distribution. The estimates reported correspond to incidence rate ratios.

	(1) Regime	(2) Flow	(3) Stock
Co-located	1.094** [4.543]	1.093** [5.406]	1.106** [5.928]
Less than 10 miles apart	1.105** [4.029]	1.064** [3.006]	1.077** [3.481]
Between 10 & 50 miles	1.188** [4.258]	1.189** [5.264]	1.203** [5.428]
Between 50 & 100 miles	1.053 [1.294]	1.077* [2.284]	1.079* [2.187]
Between 100 & 250 miles	1.123** [4.403]	1.064** [2.891]	1.091** [3.805]
Between 250 & 500 miles	1.107** [4.437]	1.043 [2.233]	1.067** [3.260]
Between 500 & 1000 miles	1.082** [3.959]	1.046** [2.709]	1.062** [3.403]
Between 1000 & 2000 miles	1.069** [3.206]	1.002 [0.124]	1.028 [1.525]
Coauthorship Effect	1.065** [3.151]	1.053** [3.193]	1.008** [3.862]
Co-located × Coauthorship Effect	0.980 [0.861]	0.965* [2.097]	0.994* [2.651]
Less than 10 miles × Coauthorship Effect	0.929** [2.727]	0.981 [1.045]	0.996† [1.650]
Between 10 & 50 miles × Coauthorship Effect	0.981 [0.432]	0.967 [1.213]	0.994 [1.459]
Between 50 & 100 miles × Coauthorship Effect	1.032 [0.688]	1.002 [0.073]	0.999 [0.286]
Between 100 & 250 miles × Coauthorship Effect	0.897** [3.558]	1.006 [0.244]	0.989** [2.666]
Between 250 & 500 miles × Coauthorship Effect	0.886** [4.396]	1.019 [0.865]	0.991** [3.016]
Between 500 & 1000 miles × Coauthorship Effect	0.926** [3.229]	0.993 [0.334]	0.993* [2.367]
Between 1000 & 2000 miles × Coauthorship Effect	0.869** [5.542]	0.970 [1.377]	0.987** [4.661]
Ln(Med. School NIH Funding)	1.047** [8.735]	1.046** [8.579]	1.046** [8.622]
Observations	570,758	570,758	570,758
Number of dyads	25,506	25,506	25,506
Log Likelihood	-268,089	-268,078	-268,100
Sum of Spillover Coeffs. = 0	45.185**	40.534**	37.246**

Notes:

(1) Absolute value of z statistics in brackets.

(2) Dependent variable is number of publications w/o the superstar, peer first or last author.

(2) All models include a set of calendar year dummies and career age dummies (coefficients not reported).

(3) †significant at 10%; *significant at 5%; **significant at 1%.

Table 16

Effect of superstar exposure on peer publication output. Conditional MLE estimates for the fixed effect Negative Binomial Model. The dyads considered are those corresponding to superstars above the the 99th prcntl. of the Cumulative NIH grant distribution. The estimates reported correspond to incidence rate ratios.

	(1)	(2)	(3)	(4)
Co-located	1.205** [20.8]		1.183** [18.3]	1.253** [21.62]
Co-located, 1981-85	1.012 [1.16]		0.985 [1.44]	1.026* [1.97]
Co-located, 1986-90	1.006 [0.63]		0.975* [2.47]	1.043** [3.09]
Co-located, 1991-95	0.880** [12.6]		0.878** [12.7]	0.863** [9.95]
Co-located, 1996-00	0.803** [20.9]		0.807** [20.1]	0.726** [17.86]
Co-located, 2001-03	0.701** [29.4]		0.702** [29.0]	0.623** [15.11]
Coauthorship, Regime		1.156** [17.2]	1.121** [13.2]	1.203** [18.44]
Coauthorship, Reg., 1981-85		1.048** [4.91]	1.063** [6.34]	1.040** [3.55]
Coauthorship, Reg., 1986-90		1.033** [3.46]	1.055** [5.60]	1.022* [1.97]
Coauthorship, Reg., 1991-95		0.867** [15.3]	0.897** [11.4]	0.846** [15.40]
Coauthorship, Reg., 1996-00		0.769** [27.3]	0.799** [22.9]	0.743** [26.69]
Coauthorship, Reg., 2001-03		0.690** [31.7]	0.720** [27.8]	0.670** [30.70]
Co-located × Coauthorship Reg.				0.791** [12.69]
Co-located × Coauthorship Reg., 1981-85				1.020 [0.89]
Co-located × Coauthorship Reg., 1986-90				1.032 [1.44]
Co-located × Coauthorship Reg., 1991-95				1.203** [8.21]
Co-located × Coauthorship Reg., 1996-00				1.359** [12.49]
Co-located × Coauthorship Reg., 2001-03				1.364** [8.68]
Ln(Med. School NIH Funding)	1.032** [17.9]	1.032** [17.7]	1.029** [16.4]	1.029** [16.26]
Observations	960,091	960,091	960,091	960,091
Number of Dyads	44,913	44,913	44,913	44,913
Log Likelihood	-1,398,755	-1,398,076	-1,397,231	-1,396,923

Notes:

- (1) Absolute value of z-statistics in brackets.
- (2) Dependent variable is number of publications w/o the superstar, peer first or last author.
- (2) All models include a set of calendar year dummies and career age dummies (coefficients not reported).
- (3) †significant at 10%; *significant at 5%; **significant at 1%.

Appendix I: Coauthor Name Matching

We eliminate all dyads in which the last year of collaboration with the superstar takes place within 5 years of the graduation year for the colleague. Such dyads do not correspond to peer relationships, since the colleague was most likely a fellow or a graduate student at the time of coauthorship. Because the roster is limited to faculty, the S/CG will not match postdoctoral fellows who fail to secure an academic position. In other words, we impose the requirement that at least some collaboration with a superstar take place after the colleague starts his/her faculty career, although the onset of collaboration could certainly predate the start of an independent faculty career. Furthermore, we impose “home-bias.” When the S/CG matches more than one faculty in the roster, but only one of these matches correspond to a co-located colleague, we eliminate the dyads corresponding to more distant colleagues. This appears warranted given the marked propensity to coauthor locally.