This PDF is a selection from a published volume from the National Bureau of Economic Research

Volume Title: The Rate and Direction of Inventive Activity Revisited

Volume Author/Editor: Josh Lerner and Scott Stern, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-47303-1; 978-0-226-47303-1 (cloth)

Volume URL: http://www.nber.org/books/lern11-1

Conference Date: September 30 - October 2, 2010

Publication Date: March 2012

Chapter Title: The Diffusion of Scientific Knowledge across Time and Space: Evidence from Professional Transitions for the Superstars of Medicine

Chapter Authors: Pierre Azoulay, Joshua S. Graff Zivin, Bhaven N. Sampat

Chapter URL: http://www.nber.org/chapters/c12350

Chapter pages in book: (p. 107 - 155)

2

The Diffusion of Scientific Knowledge across Time and Space

Evidence from Professional Transitions for the Superstars of Medicine

Pierre Azoulay, Joshua S. Graff Zivin,
and Bhaven N. Sampat

If Whitehead's characterization of modern philosophy as "a series of footnotes to Plato" is perhaps a slight exaggeration, then the claim that contemporary scholarship on the economics of innovation is largely an extension of themes laid out in the 1962 *Rate and Direction* volume is no more of one. The contributors' prescience about the potential economic importance of academic science is particularly impressive, since at the time the conference was held (in 1960), the post-Sputnik transformation of the academic enterprise into the behemoth we know today had only just begun. The volume laid out a belief that basic research was important for innovation, marshalling theory, case studies, and data to support the assertions about the economic payoffs from basic research made in Vannevar Bush's "Science: the Endless Frontier" just fifteen years before. However, the conference volume was very much a call for more research, emphasizing the need for more data, and deeper understanding. In this chapter we attempt to rise to this challenge by examining the diffusion of knowledge across time and space within the life sciences. This endeavor remains as important at the beginning of this century as it was in the middle of the last one, given perennial calls for

Pierre Azoulay is associate professor at the Sloan School of Management, Massachusetts Institute of Technology, and a faculty research fellow of the National Bureau of Economic Research. Joshua S. Graff Zivin is associate professor of economics at the School of International Relations and Pacific Studies at the University of California, San Diego, and a research associate of the National Bureau of Economic Research. Bhaven N. Sampat is assistant professor in the Department of Health Policy and Management at the Mailman School of Public Health at Columbia University.

Send correspondence to pazoulay@mit.edu. We gratefully acknowledge the financial support of the National Science Foundation through its SciSIP Program (Award SBE-0738142). We thank the conference audience, our discussant Adam Jaffe, as well as Scott Stern and Manuel Trajtenberg, for useful comments and suggestions. The usual disclaimer applies.

justification of substantial public funds for biomedical research (especially during periods of fiscal austerity) and current attempts to ground "science of science policy" on stronger theoretical and empirical footing (Marburger 2005).

Our analyses share a main motivation with the original conference, to understand how nonmarket controls and incentives at universities operate, and affect innovation. (The subtitle of the *Rate and Direction* volume is, after all, "Economic and Social Factors.") Over the past five decades, a voluminous literature on the workings of academic science has emerged in economics, sociology, and other disciplines (Stephan 2010; Dasgupta and David 1994; Merton 1973). Similarly, the rise of endogenous growth theory (Romer 1990; Aghion and Howitt 1992) and its emphasis on spillovers has focused attention on how knowledge flows across individuals, locations, and institutional settings. A particular focus has been on the extent to which knowledge flows are geographically localized (see, *inter alia,* Jaffe, Trajtenberg, and Hendeson [1993]; Thompson and Fox-Kean [2005]; and the response by Henderson, Jaffe, and Trajtenberg [2005]). Location was not a central concern in the 1962 volume (with the exception of the chapter by Wilbur Thompson). However, it has become an important policy issue since. The extent to which knowledge flows are geographically mediated is relevant to local and national policymakers, in deciding whether the benefits of the research they fund will accrue to those that fund it, or diffuse more generally.

A second theme in the volume is the difficulty in measuring economic activity. Several chapters explored the utility of patent data as indicators of innovation, and also emphasized that patent data alone may paint a distorted picture of the rate and direction of innovation (Kuznets 1962). Measuring knowledge flows is perhaps even more difficult than measuring innovation, since these flows leave few footprints (Krugman 1991). Nonetheless, a long literature in sociology and bibliometrics has attempted to measure knowledge flows among academics, using publication-to-publication citations. More recently, economists have employed patent-to-patent citations to examine knowledge flows from academics to industry (Jaffe and Trajtenberg 1999). A few papers (Branstetter 2005; Belenzon and Schankerman 2010) also use patent-publication citations.

Our study joins a small but distinguished literature relating patterns of citations to individual mobility. Almeida and Kogut (1999) use a sample of highly cited semiconductor patents, and information on citations to these patents (and a control sample of other patents in the same class as citing patents) to examine the extent and determinants of citation localization in this industry. They also identify the set of inventors on these patents who had moved previously, constructing career paths using patent records. The authors use these data to distinguish between regions with high intra- and interregional mobility, and find that patents from regions with high intra-

regional mobility are more likely to have citations that are local, and patents from regions with high interregional mobility are less likely to have local citations.[1]

Using a similar research design, but one more closely related to our own, Agrawal, Cockburn, and McHale (2006) examine all US patents applied for in 1989 and 1990 by movers, operationalized as individuals with the same names who had previously patented in the same patent class. Their analysis shows that the citations to postmove patents emanating from the inventors' prior location are disproportionately high, estimating that 50 percent more of the citations to postmove patents come from the prior location than would have if the inventor had not previously lived there. Since this citation premium to postmove patents is unlikely to reflect low communication costs or direct interaction (variables often invoked in explaining why geography matters), they interpret these results as evidence of the enduring importance of social relationships.

Our analysis also departs from these previous analyses in important ways: we identify movers from scientists' vitae (rather than patent data); we examine cited and citing knowledge longitudinally, exploiting detailed information on the timing of the move; and we look at three distinct measures of knowledge flows. The use of multiple indicators allows us to assess not only whether knowledge flows from academe are geographically mediated, but also to probe some of the mechanisms that might underlie this relationship—in short, to deepen our understanding of knowledge diffusion and its implications for the level and rate of technological innovation within the economy.

We examine these issues using a novel identification strategy that exploits labor mobility in a sample of 9,483 elite academic life scientists to examine the impact of moving on the citation trajectories associated with individual articles (respectively patents) published (respectively granted) *before* the scientist moved to a new institution. This longitudinal contrast purges our estimates of most sources of omitted variable bias that can plague cross-sectional comparisons. However, the timing of mobility itself could be endogenous. To address this concern, we pair each moving scientist/article dyad (respectively scientist/patent dyad) with a carefully chosen control article or patent associated with a scientist who does not transition to a new position. In addition to providing a very close match based on time-invariant characteristics, these controls also share very similar citation trends prior to the mobility event. By analyzing the data at the matched-pair level of analysis, this simple difference-in-difference framework provides a flexible and nonparametric methodology to evaluate the effects of labor mobility on knowledge flows.

1. In some analyses, Almeida and Kogut also explore individual (rather than regional) level mobility, finding that inventors who move within a region tend to have citations that are geographically local.

Indeed, conditional on the assumption that the matching algorithm we employ successfully pairs articles and patents of comparable quality, we are able to present the findings in a straightforward, graphical form.

The results reveal a nuanced story. We find that article-to-article citations from a scientist's origin location are barely affected by their departure. In contrast, patent-to-article citations, and especially patent-to-patent citations, decline at the origin location following a superstar's departure, suggesting that spillovers from academia to industry are not completely disembodied. We also find that article-to-article citations from a scientist's destination location markedly increase after they move. To the extent that academic scientists do not internalize the effect of their location decisions on the circulation of ideas, our results raise the intriguing possibility that barriers to labor mobility in academic science limit the recombination of individual bits of knowledge, resulting in a suboptimal rate of scientific exploration.

The chapter proceeds as follows. The next section discusses the construction of our multilevel panel data set and presents relevant descriptive statistics. Section 2.2 discusses our econometric approach and identification strategy. Section 2.3 reports the results. The final section includes a discussion of policy implications, caveats, and directions for future research.

2.1 Data and Sample Characteristics

The setting for our empirical work is the academic life sciences. This sector is an important one to study for several reasons. First, there are large public subsidies for biomedical research in the United States. With an annual budget of $29.5 billion in 2008, support for the National Institutes of Health (NIH) dwarfs that of other national funding agencies in developed countries (Cech 2005). Deepening our understanding of how the knowledge generated by these expenditures diffuses across time, space, and institutional settings will allow us to better assess the return to these public investments.

Second, technological change has been enormously important in the growth of the health care economy, which accounts for roughly 15 percent of US gross domestic product (GDP). Much biomedical innovation is science-based (Henderson, Orsenigo, and Pisano 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, Darby, and Brewer 1998).

Lastly, the existence of geographic research clusters in the life sciences has been extensively documented, raising the possibility that scientific knowledge diffuses only slowly and with a lag from areas richly endowed with academic research institutions to others. To the extent that scientist labor mobility is needed to support the circulation of ideas to the periphery, a

dearth of mobility events might be one of the centripetal forces leading to the persistence of such clusters.

In the next section, we provide a detailed description of the process through which the matched scientist/article (resp. scientist/patent) data set used in the econometric analysis was assembled. We begin by describing the criteria used to select our sample of superstar life scientists, along with basic demographic information. Next, we explore the prevalence and characteristics of mobility events; the set of products (i.e., journal articles and patents) generated by these elite scientists along with the citations they accrue. Finally, we discuss the matching procedure implemented to identify control articles and patents associated with scientists who do not change their location.

2.1.1 Superstar Sample

Our basic approach is to rely on professional transitions in a sample of "superstar" scientists in the United States to estimate the extent to which citation flows to individual pieces of knowledge are constrained by their producers' geographic location.

The study's focus on the scientific elite can be justified both on substantive and pragmatic grounds. The distribution of publications, funding, and citations at the individual level is extremely skewed (Lotka 1926; de Solla Price 1963) and only a tiny minority of scientists contribute through their published research to the advancement of science (Cole and Cole 1972). Furthermore, analyzing the determinants of citations flowing to the ideas of elite scientists is arguably more interesting than conducting the same exercise for a sample of less distinguished scientists, since superstars presumably produce knowledge that is more important to diffuse.

From a practical standpoint, it is also more feasible to trace back the careers of eminent scientists than to perform a similar exercise for less eminent ones. We began by delineating a set of 10,450 "elite" life scientists (roughly 5 percent of the entire relevant labor market) who are so classified if they satisfy at least one of the following criteria for cumulative scientific achievement: they are (a) highly funded scientists; (b) highly cited scientists; (c) top patenters; or (d) members of the National Academy of Sciences.

These four criteria naturally select seasoned scientists, since they correspond to extraordinary achievement over an entire scientific career. We combine these measures with three others that capture individuals who show great promise at the early and middle stages of their scientific careers, whether or not these episodes of productivity endure for long periods of time: scientists who are (e) NIH MERIT awardees; (f) Howard Hughes Medical Investigators; or (g) early career prize winners. Appendix A provides additional details regarding these seven indices of "superstardom."

We trace back these scientists' careers from the time they obtained their first position as independent investigators (typically after a postdoctoral

fellowship) until 2006. We do so through a combination of curricula vitæ, NIH biosketches, *Who's Who* profiles, accolades/obituaries in medical journals, National Academy of Sciences biographical memoirs, and Google searches. For each one of these individuals, we record employment history, degree held, date of degree, gender, and up to three departmental affiliations. We also cross-reference the list with alternative measures of scientific eminence. For example, the elite subsample contains every US-based Nobel Prize winner in Medicine and Physiology since 1975, and a plurality of the Nobel Prize winners in Chemistry over the same time period.[2]

The 9,483 scientists who are the focus of this chapter constitute a subset of this larger pool of 10,450. We impose several additional criteria to derive the final list. First, we eliminate from the sample scientists who transition from academic positions to jobs in industry; second, we eliminate scientists who move to foreign institutions, since we have less ability to track knowledge flows to these locations; third, we eliminate scientists who move twice in quick succession, since these cases make it difficult to assign to these individuals unique origin and destination locations. Finally, we eliminate scientists who moved to new institutions prior to 1975, the beginning of our observation window.

Turning to patterns of labor mobility, we find that 2,894 scientists (30 percent) in the sample transitioned between two academic institutions between 1975 and 2004. Our mobility data is tabulated precisely from biographical records, rather than inferred from affiliation information in papers or patents (cf., Almeida and Kogut 1999; Fallick, Fleischmann, and Rebitzer 2006; Marx, Strumsky, and Fleming 2009). In particular, we observe the exact timing of professional transitions even in the cases in which a scientist has ceased to be active in research; for example, because she or he has moved into an administrative position. Because the overwhelming majority of mobility events take place in the summer, we adopt the following convention: a scientist is said to move from institution A to institution B in year t whenever the actual timing of his or her move coincided with the summer of year $t - 1$. Incorporating a lag is necessary, since life scientists need to move entire laboratories rather than simply books and computer equipment. Anecdotal evidence suggests that mobility disrupts the pace of these scientists' research activities, if only temporarily.

We focus on transitions between distant institutions; that is, those separated by at least fifty miles. This limitation can be justified on both substantive and pragmatic grounds. First, many of the social impediments to labor mobility (such as dual-career concerns or disruption in the lives of these scientists' children) are less salient for professional transitions that do not compel an individual to change his place of residence. Second, our ability

2. Though we apply the term of superstar to the entire group, there is substantial heterogeneity in intellectual stature within the elite sample (see table 2.1).

Table 2.1 **Superstar scientists' cumulative output by 2006 or career end**

	Mean	Median	Std. dev.	Min.	Max.
Stayers (N = 6,589)					
NIH funding	$17,491,538	$11,261,535	$25,598,484	$0	$588,753,152
Publications	171	142	125	2	1,167
Patents	3.29	0	9.79	0	258
Paper cites [to papers]	10,639	7,332	11,248	15	139,872
Patent cites [to papers]	117	67	166	0	1,728
Patent cites [to patents]	91	19	240	0	5,596
Movers (N = 2,894)					
NIH funding	$16,256,723	$12,373,582	$16,243,082	$0	$195,611,552
Publications	174	144	121	1	1,631
Patents	3.15	0	8.30	0	117
Paper cites [to papers]	10,878	7,455	10,533	2	83,301
Patent cites [to papers]	120	69	159	0	1,821
Patent cites [to patents]	67	15	164	0	2,079

Notes: Sample consists of 9,483 elite academic life scientists. Movement is defined by a change in academic institution with at least fifty miles separating origin and destination.

Table 2.2 **Demographic characteristics**

	Degree year	Female	MD	PhD	MD/PhD
Stayers (N = 6,589)	1970.3	0.14	0.33	0.58	0.10
Movers (N = 2,894)	1972.7	0.14	0.30	0.61	0.10
Total (N = 9,483)	1971.0	0.14	0.32	0.59	0.10

to assign precisely the institutional affiliation of citing authors and inventors is limited. Therefore, we define an elite scientist's location by drawing a twenty-five-mile radius circle centered around the middle of the zip code in which his employer is located. Combined with our emphasis on moves between institutions separated by at least fifty miles, this ensures that origin and destination locations never overlap in the subsample of scientists that move.

Tables 2.1 and 2.2 provide descriptive statistics for the superstar sample. The gender composition of the sample is heavily skewed, no doubt because our metrics of superstardom favor more seasoned scientists, who came of age before female scientists had made significant inroads in the professoriate. The average degree year is 1971, and MDs account for a third of the sample. On the output side, the stars received an average of roughly seventeen million dollars in NIH grants and published 172 papers that garnered close to 11,000 citations as of early 2008. The number of patents per scientist is considerably smaller, and close to 40 percent of the sample scientists have no patent at all. While patents and papers can each appear as prior art cited in subsequent patents, the number of such citations is quite modest compared

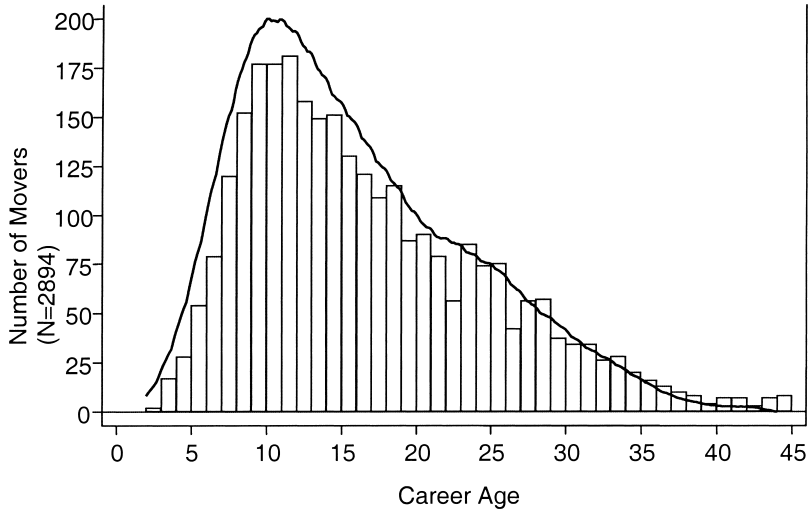

Fig. 2.1 Career age at time of move
Note: Nine observations between 46 and 54 years omitted.

to the number of article-to-article citations.[3] Achievement and demographic characteristics appear broadly similar between "moving" (i.e., treated) and "staying" (i.e., control) stars.

Figure 2.1 displays the distribution of career age at the time of move in the subsample of movers. The likelihood of a mobility event peaks at about twelve years (career age is measured as the number of years that elapsed since the receipt of one's highest degree). Figure 2.2 displays the distribution of distance moved, conditional on a move. That the shape of this distribution is strongly bimodal is not surprising, given the existence of life sciences research clusters on both coasts of the United States. Finally, figure 2.3 examines whether our elite scientists systematically drift from areas rich in the relevant type of intellectual capital to areas less well endowed (or vice versa). We compute total NIH funding flowing to scientists' origin and destination areas (panel A) and repeat the same exercise with the number of patents issued to inventors located in these same areas (panel B). While not symmetric in a strict statistical sense, these histograms make clear that most of the transitions in the sample involve relatively little difference in the resource endowments of the relevant locations, while a few are big moves in the sense of taking a scientist away from a less prestigious institution into a more intellectually vibrant climate (or vice versa).

3. Nonetheless, it is striking that the mean number of citations to these scientists' papers is larger than those to their patents. This difference (even more pronounced when considering the medians) is consistent with the results of Cohen, Nelson, and Walsh (2002), who find that the bulk of knowledge flows from academe to industry occur via open science channels.

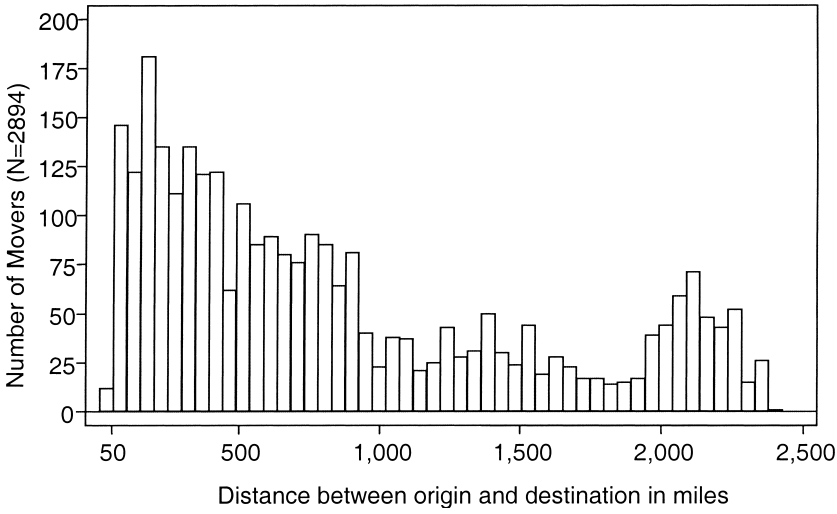

Fig. 2.2 Distance moved
Note: Five observations between 2,500 and 4,500 miles omitted.

2.1.2 Matching Scientists with Their Output

The second step in the construction of our data set is to link scientists with the knowledge they generate in tangible form, namely journal articles and patents. A useful metaphor for this exercise is that of linking producers with their products. Past scholarship in the field of the economics of science has generated numerous studies that rely on variation between individual producers, either cross-sectionally (Zucker, Darby, and Brewer 1998), or over time (Azoulay, Ding, and Stuart 2009; Azoulay, Graff Zivin, and Wang 2010), while paying scant attention to the detailed characteristics of the products involved. Conversely, a more recent and vibrant strand of the literature has exploited the availability of citations to individual products over time, for the most part abstracting away from the characteristics of their producers (Furman and Stern, forthcoming; Aghion et al. 2009).[4]

A major innovation in our study is to link detailed producer and product characteristics to create a multilevel panel data set.[5] Social scientists face difficult practical constraints when attempting to attribute individual products to particular producers. When the products involved are journal articles, there are thorny issues of name uniqueness: common names make it difficult to distinguish between scientists, and even scientists with relatively

4. In what follows, the use of the word "product" will also be useful whenever we want to refer to the output of our elite scientists in a generic way so that our statements apply equally well to journal articles and to patented inventions.

5. Recent efforts along the same line include Agarwal and Singh (forthcoming) and Azoulay, Stuart and Wang (2010).

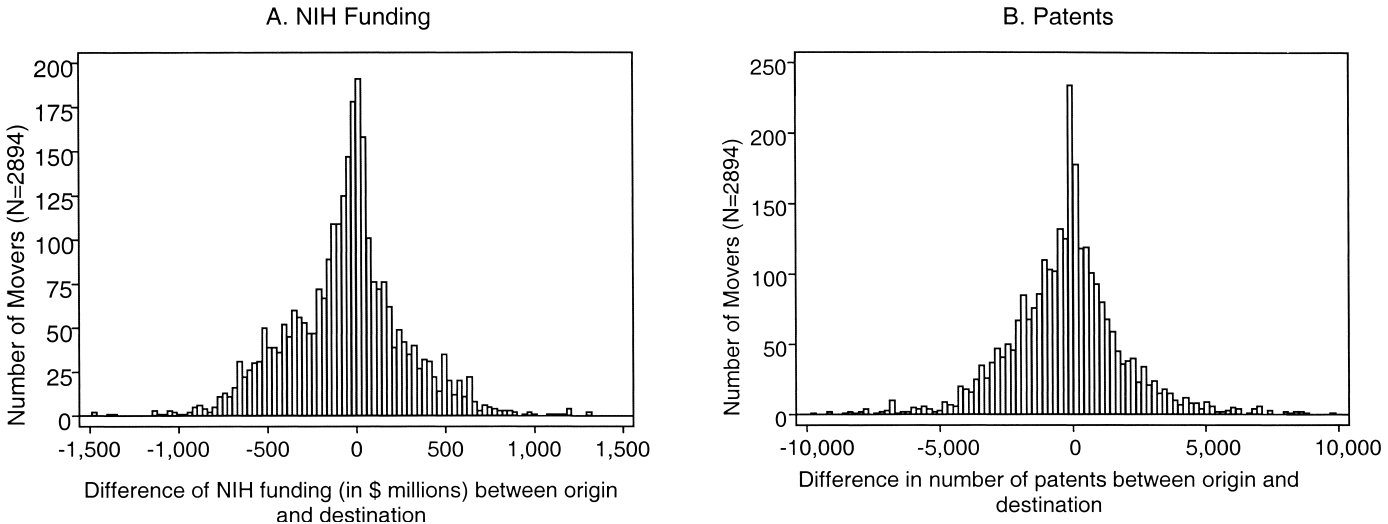

Fig. 2.3 Similarity between origin and destination locations

Note: The preceding histograms map the extent to which professional transitions in our sample take elite scientists to destinations that differ from, or are similar to, the areas from which they originate, along two dimensions: NIH funding accruing to institutions located in 25-mile radius centered on the origin and destination institution (panel A); and the number of patents applied for by inventors located in 25-mile radius centered on the origin and destination institution (panel B).

rare names sometimes are inconsistent in their use of publication names. By adopting the labor-intensive practice of designing customized search queries for each scientist in the sample, we ensured the accuracy of their bibliomes. Further details on the linking process are provided in appendix B. Linking scientists with their patented inventions is considerably easier, since the United States Patent and Trademark Office (USPTO) data records inventors' full names (as opposed to first and middle initials), and we can further make use of assignee information to distinguish between the patents of inventors with frequent names. Further details on the linking process are provided in appendix C.

We select from these publication and patent data to construct the final sample. For journal articles, we eliminate from the consideration set letters, comments, and editorials. Second, we eliminate all articles published eleven or more years prior to the date of the earliest move in the sample (1976); similarly, we eliminate all articles published in 2004 (the latest move year we observe) or in subsequent years. Third, we delete from the sample all articles published by moving scientists after they moved. We proceed similarly for patents. To account for potential truncation, we assume an average grant lag of three years, and we ignore all patents applied for after the year 2001.

2.1.3 Three Measures of Knowledge Flows

As noted by many authors, beginning with the seminal work of Jaffe, Trajtenberg, and Henderson (1993), knowledge flows sometimes leave a paper trail, in the form of citations in either patents, or journal articles. The innovation in the present study is that we present evidence pertaining to three distinct measures of knowledge flows: citations to articles authored by our elite scientists in the open science literature; citations to articles authored by our elite scientists listed in the prior art section of patents issued by the US Patent and Trademark Office (USPTO); and citations to patents granted to our elite scientists in patents subsequently granted to other inventors by the USPTO. Each of these measures exhibits a particular set of strengths and weaknesses. We describe them in turn.

Patent-to-Patent Citations. The bulk of the voluminous research on knowledge spillovers has relied on patent citations in other patents to infer patterns of knowledge diffusion (cf. Jaffe and Trajtenberg 1999). The difficulties involved in interpreting these citations as evidence of knowledge flows—mostly because of the high share of citations added by examiners, rather than assignees—have been explored in detail (Alcácer and Gittelman 2006; Alcácer, Gittelman, and Sampat 2009) and need not be repeated here. Despite these acknowledged problems, survey results confirm that roughly 50 percent of patent-to-patent citations represent some sort of knowledge flow (Jaffe, Trajtenberg, and Fogarty 2002). Moreover, the prevalence of examiner-added citations is much smaller in the life sciences than in other fields (Sampat 2010).

Article-to-Article Citations. Beyond the low "signal-to-noise" ratio associated with patent citations, a more serious limitation for our purposes is that the bulk of the output of the academics we study is in publications, rather than patents. Fully 60 percent of the scientists in our superstar sample never apply for a patent, and the great majority of those who do patent have only one or two inventions to their credit. Therefore, we also collect the number of citations in subsequent journal articles that flow to each of the papers generated by our superstars, over time. A great advantage of these citations is that they are numerous, making it possible to parse these data in ever finer slices to tease out the underlying mechanisms that support the diffusion of scientific knowledge. Their main drawback is that 95 percent of citations flowing to the articles in our sample come from other academics. These data are therefore less useful to track the flow of ideas across the boundary between academia and for-profit firms.

Patent-to-Article Citations. These limitations lead us to introduce a novel measure of knowledge flows, namely, references to the open science literature found in the nonpatent prior art section of patents granted by the USPTO. This is appealing both because publications rather than patents are the main output of scientific researchers (Agrawal and Henderson 2002), but also because the vast majority of patent-to-paper citations, over 90 percent, come from applicants rather than examiners, and are thus more plausible indicators of real knowledge flows than patent-to-patent citations (Lemley and Sampat 2010). Another advantage of these data comes from the greater diversity of citing institution types, relative to the patterns exhibited by the more traditional data sources mentioned earlier. In previous work, systematic analyses of these nonpatent references has been limited, since they are free-form text and difficult to link to other data. Our work relies on a novel match between nonpatent references and biomedical articles indexed in PubMed, described in detail in appendix D. While programming improvements and computing speed have enabled us to mine this source of data, only 12 percent of the published output of the scientists in our sample is ever cited in patents. For this reason, the bulk of our analyses will focus on citation flows inferred from article-to-article citations.

After collecting the citation data, we further process it in order to make it amenable to statistical analysis. First, we eliminate all self-citations since these do not correspond to knowledge flows in the traditional sense.[6] Second, we parse the address fields in both the citing patents and citing publications to associate each citing product with a set of zip codes (for US addresses) or country names (for foreign addresses). Third, we parse the citing assignee names and citing institution names and tag these fields with an

6. In the case of patents, we infer self-citation from overlap between the names of inventors in the cited and citing patents, rather than overlap in assignee names.

indicator variable denoting an industrial affiliation, making use of suffixes such as Inc., Corp., Ltd. (or their international variants). In a final step, we aggregate the data from the cited product-citing product pair level up to the cited product-year level of analysis. In other words, we can track the flow of citations from birth to 2006 for each producer/product tuple in the sample.[7]

We can further separate those citations that accrue to a scientist's origin location, to his or her destination, or to all other locations. A complication arises because it is not clear what destination means for the sample of superstars who do not transition to a new location. Ideally, we would select as a counterfactual location the institution that provides the highest degree of fit for these scientists outside of their actual home institution. In practice, it is very difficult to model the determinants of fit, and we select a location at random from the set of locations that moving scientists transition to, provided they are separated by at least fifty miles from the stayers' actual locations.

2.1.4 From Control Producers to Control Products: A Nonparametric Matching Procedure

A perennial challenge in the literature on the localization of knowledge flows is whether citing and cited producers' locations can be credibly assumed to be exogenous. Henderson, Jaffe, and Trajtenberg (2005) describe this thorny issue:

> Professor Robert Langer of MIT, for example, is one of the world's leading experts in tissue engineering, and is the author of over 120 patents in the area. A large fraction of the citations to these patents are geographically localized. Are they local just because the authors of the citing patents lived in the same city and hence were more likely to learn about Langer's work (i.e., knowledge spillovers)? Or because Boston is one of the world's centers for tissue engineering, and so people working in the area are disproportionately likely to live in Boston (i.e., geographic collocation due to other common factors)? Or perhaps it is the case that Boston is one of the world's centers for tissue engineering precisely because firms locate in the area in order to be able to take advantage of spillovers from people like Robert Langer?

Previous scholars faced severe data constraints in their attempts to divine whether a particular citation would have taken place, if contrary to the fact, either the citing or the cited producer had been located elsewhere (Jaffe, Trajtenberg, and Henderson 1993; Jaffe and Trajtenberg 1999). In this study,

7. Since the latest year in which a scientist moves is 2004, and the latest product vintage we include in the sample is 2003, the postmove observation period will always extend for a minimum of three years.

we can relax these constraints and design more credible counterfactuals, since we can "unbundle" producers from their products, and we are able to observe two different locations for a significant subset of the producers in the sample.

Yet, relying on labor mobility to generate variation in the geographic distance separating the source and potential recipients of knowledge is not a panacea for two reasons. First, producer mobility might influence the quality of the underlying products, for instance, because the scientist finds himself or herself located in an institution for which she or he is a better match. We deal with the threat of unobserved heterogeneity of this type by narrowing our focus to products generated by scientists *prior* to their move. It is difficult to imagine a mechanism through which the quality of these products could have been affected by the characteristics of the destination location.

Second, it is possible that job transitions for academic scientists are partly driven by expectations of interactions with academic peers in their home institution, or with the local industrial base. To generate a set of estimates that can be given a causal interpretation, we create the matched sample of "staying" producers described earlier, which we link to their products following the exact same techniques.

Coarse Exact Matching Procedure. We design a procedure to cull from the universe of products associated with "control" producers (i.e., scientists who do not change locations) a subset that provides a very close match with the products of "treated" producers (i.e., those scientists who do move to another institution at some point during the observation period). The goal of the construction of this matched sample is to create for the nonmovers a counterfactual set of products that mimic the citation trajectories associated with movers' papers and patents.

What makes a good control? Control and treated products should be well matched on time-invariant characteristics that have an important impact on the magnitude of citation flows. For journal articles, such characteristics might include the journal in which the article appeared, the exact time of publication, the number of scientists on the article's authorship list, and so forth. For patents, finding a control such that application year, issue year, number of inventors, assignee type, and patent classes/subclasses coincide would be valuable. More importantly, there should be no differential citation trends that affect treated products, relative to control products, in the period that precedes the move. Finally, in an ideal world, the match would operate at the producer/product pair level, such that focal producer characteristics (age, gender, and eminence) would also be comparable between treated and control observations.

In practice, identifying close matches is difficult. Because we are interested in the fate of individual products, but the shock we observe (mobility) operates at the scientist-level of analysis, semiparametric matching techniques

(such as the propensity score and its variants) are of limited use in our context. We propose instead a nonparametric matching approach, a so-called "coarse exact matching" (CEM) procedure (Blackwell et al. 2009).

The selection of controls proceeds in a series of sequential steps. The first task is to select a relatively small set of covariates on which we would like to guarantee balance between the treatment and control group. The second step is to create a large number of strata to cover the entire support of the joint distribution of the covariates selected in the previous step. Next, each observation is allocated to a unique strata; any strata that either has no product associated with a mover, or that has less than five potential control products, is then dropped from the data. In a fourth and final step, we select in each strata a unique control product such that the sum of squared differences in citation flows between the treated and control product from the year of publication/issue up to the year preceding the move year is minimized. We break ties at random when there are several candidate products that minimize this distance metric.

Internal versus External Validity. The procedure is coarse because we do not attempt to precisely match on covariate values; rather, we coarsen the support of the joint distribution of the covariates into a finite number of strata, and we match a treated observation if and only if a control observation can be recruited from this strata. An important advantage of CEM is that the analyst can guarantee the degree of covariate balance ex ante, but this comes at a cost: the more fine-grained the partition of the support for the joint distribution (i.e., the higher the number of strata), the larger the number of unmatched treated observations. In general, the analyst must trade off the quality of the matches with external validity: the longer the list of matching covariates, the more difficult it is to identify an "identical twin" for each article or patent in the treatment group.

We illustrate the essence of the matching procedure in figures 2.4 (for articles) and 2.5 (for patents). Implementation details can be found in appendix E. In the case of article-to-article citations, we start from a universe of 40,023 papers corresponding to the published output of movers in the 10 years that precede their change in location. We match 10,249 out of these 40,023 tuples (25.61 percent). This relatively low match rate is not surprising. Nonparametric matching procedures such as CEM are prone to a version of the "curse of dimensionality" whereby the proportion of matched units decreases rapidly with the number of strata. For instance, requiring a match on an additional indicator variable (e.g., matching on focal scientist gender in addition to the covariates mentioned earlier) would result in a match rate of about 10 percent. Conversely, failing to impose that control and treated articles are drawn from the same scientific journal would increase the match rate to 70 percent, but doing so might threaten the internal validity of our empirical exercise. In the case of article-to-patent citations, we match 2,435 articles out of a potential 6,492 (37.51 percent). In the case of

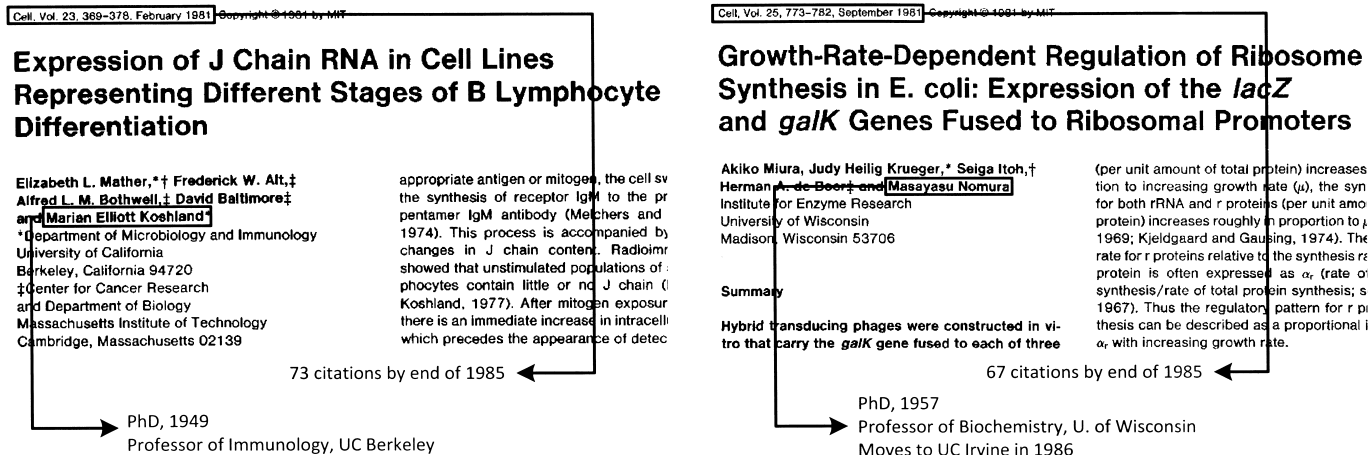

Fig. 2.4 Article-level match

Note: The two articles illustrate the essence of the coarse exact matching procedure. These two articles appeared in the journal *Cell* in 1981. They received a very similar number of citations up to 1985: 73 citations for Mather et al.; 67 citations for Miura et al. Masayasu Nomura, the PI on the article on the right-hand side, moved from the University of Wisconsin to UC Irvine in 1986.

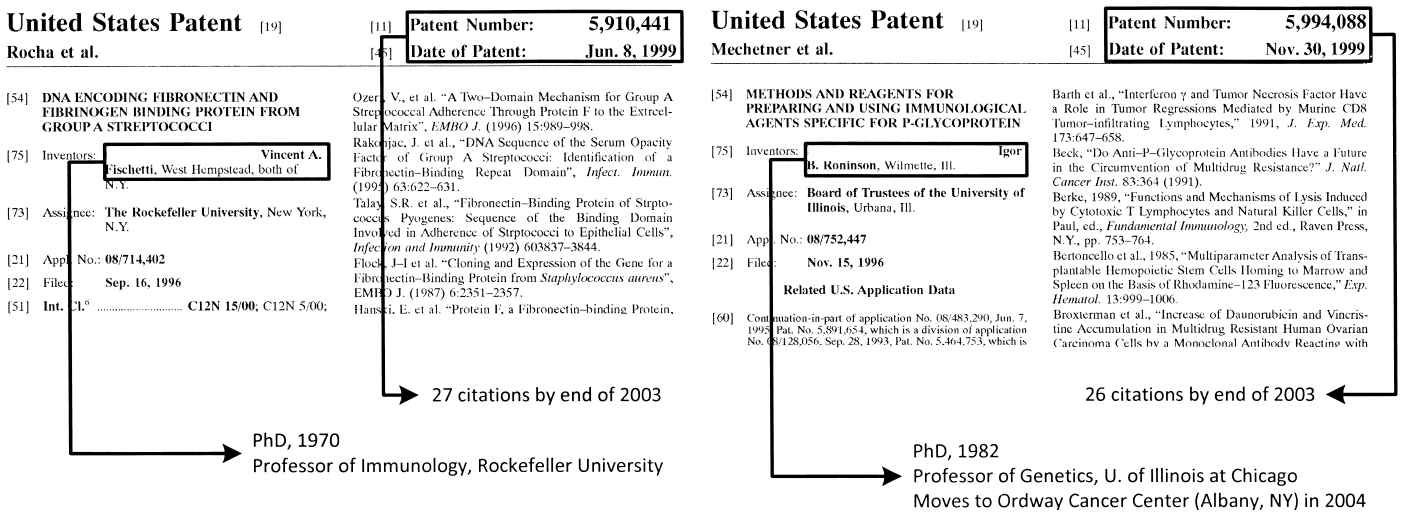

Fig. 2.5 Patent-level match

Note: The two patents illustrate the essence of the coarse exact matching procedure. These two patents were issued in 1999 and applied for in 1996. They both belong to patent class 435 (Molecular Biology & Microbiology). They received a very similar number of citations up to 2003: 27 citations for Fischetti's (Richa et al.); 26 citations for Roninson's (Mechetner et al.). Igor Roninson—the focal inventor on patent 5,994,088—moved from the University of Illinois at Chicago to the Ordway Cancer Center in Albany, New York, in 2004.

patent-to-patent citations, the match rate is higher still: 41.36 percent (1,417 matched patents out of a potential 3,426 matches).

Descriptive Statistics. We present univariate statistics at baseline—that is, in the year preceding the (possibly counterfactual) move year—for the matched product data sets in tables 2.3, 2.4, and 2.5. Examining the raw data across these three panel data sets, a number of stylized facts emerge.

First, in all three cases, the match is "product-centric" rather than "producer-centric." That is, product-level attributes exhibit a high level of

Table 2.3 **Article-to-article citation flows: Descriptive statistics ($n = 2 \times 10{,}249$), articles published *before* the move**

	Mean	Median	Std. dev.	Min.	Max.
Journal Articles by Stayers					
Number of authors	4.463	4	2.980	1	129
Focal author is last	0.653	1	0.476	0	1
Article age at baseline	2.483	2	2.055	1	10
Focal author gender	0.098	0	0.297	0	1
Focal author graduation year	1967.491	1968	10.893	1931	2001
Article baseline stock of article citations	27.666	3	66.750	0	2399
Article baseline stock of article citations from industry	1.023	0	3.731	0	135
Article baseline stock of article citations at origin	1.952	0	6.550	0	135
Article baseline stock of article citations at destination	0.355	0	2.210	0	80
Journal Articles by Movers					
Number of authors	4.489	4	3.238	1	180
Focal author is last	0.653	1	0.476	0	1
Article age at baseline	2.483	2	2.055	1	10
Focal author gender	0.084	0	0.277	0	1
Focal author graduation year	1972.603	1973	9.289	1940	1997
Article baseline stock of citations	27.824	3	63.855	0	1226
Article baseline stock of article citations from industry	1.036	0	3.477	0	103
Article baseline stock of article citations at origin	1.834	0	6.284	0	149
Article baseline stock of article citations at destination	0.624	0	3.249	0	131

Notes: The match is article centric; that is, the control article is always chosen from the same journal in the same publication year. The control article is coarsely matched on the number of authors (exact match for one, two, and three authors; four or five authors; between six and nine authors; and more than nine authors). We also match on focal scientist's position in the authorship roster (first author; last author; middle author). For articles published one year before appointment, we also match on the month of publication. For articles published two years before appointment, we also match on the quarter of publication. In addition, the articles in the control and treatment groups are matched on article citation dynamics up to the year before the (possibly counterfactual) transition year. The cost of a very close, nonparametric match on article characteristics is that author characteristics do not match closely. Imposing a close match on focal scientist age, gender, and overall productivity at baseline would result in a match rate which is unacceptably low.

Table 2.4 **Patent-to-article citation flows: Descriptive statistics ($n = 2 \times 2{,}435$), articles published *before* the move**

	Mean	Median	Std. dev.	Min.	Max.
Journal articles by stayers					
Number of authors	5.062	5	2.596	1	38
Focal author is last	0.598	1	0.490	0	1
Article age at baseline	3.118	2	2.333	1	10
Focal author gender	0.083	0	0.276	0	1
Focal author graduation year	1965.539	1967	12.022	1931	1999
Article baseline stock of patent citations	0.499	0	1.649	0	29
Article baseline stock of patent citations from industry	0.352	0	1.375	0	24
Article baseline stock of patent citations at origin	0.040	0	0.449	0	16
Article baseline stock of patent citations at destination	0.012	0	0.306	0	14
Journal articles by movers					
Number of authors	5.049	5	2.433	1	26
Focal author is last	0.598	1	0.490	0	1
Article age at baseline	3.118	2	2.333	1	10
Focal author gender	0.086	0	0.281	0	1
Focal author graduation year	1974.161	1975	8.709	1940	1995
Article baseline stock of patent citations	0.540	0	1.889	0	46
Article baseline stock of patent citations from industry	0.367	0	1.652	0	46
Article baseline stock of patent citations at origin	0.029	0	0.284	0	6
Article baseline stock of patent citations at destination	0.019	0	0.236	0	7

Notes: The match is article centric; that is, the control article is always chosen from the same journal in the same publication year. The control article is coarsely matched on the number of authors (exact match for one, two, and three authors; four or five authors; between six and nine authors; and more than nine authors). We also match on focal scientist's position in the authorship roster (first author; last author; middle author). For articles published one year before appointment, we also match on the month of publication. For articles published two years before appointment, we also match on the quarter of publication. In addition, the articles in the control and treatment groups are matched on patent citation dynamics up to the year before the (possibly counterfactual) transition year. The cost of a very close, nonparametric match on article characteristics is that author characteristics do not match closely. Imposing a close match on focal scientist age, gender, and overall productivity at baseline would result in a match rate which is unacceptably low.

covariate balance between treated and control products, whether these characteristics are time-invariant (such as number of authors or focal scientist position on the authorship roster) or time-varying (such as the stock of overall citations, whose distributions we display graphically in figure 2.6). In contrast, producer characteristics do not match as well, as can be seen by examining the distribution of covariates such as degree year or gender.

Second, most citations do not accrue in the areas corresponding to these

Table 2.5 **Patent-to-patent citation flows: Descriptive statistics ($n = 2 \times 1{,}417$), patents issued *before* the move**

	Mean	Median	Std. dev.	Min.	Max.
Patents by stayers					
Patent age at baseline	4.579	4	2.610	1	10
Focal author gender	0.056	0	0.231	0	1
Focal author graduation year	1969.762	1970	10.806	1932	1996
Patent baseline stock of patent citations	7.076	1	14.770	0	135
Patent baseline stock of patent citations from industry	5.880	0	12.830	0	98
Patent baseline stock of patent citations at origin	0.563	0	2.899	0	48
Patent baseline stock of patent citations at destination	0.167	0	1.439	0	37
Patents by movers					
Patent age at baseline	4.579	4	2.610	1	10
Focal author gender	0.047	0	0.212	0	1
Focal author graduation year	1976.711	1978	8.678	1950	1996
Patent baseline stock of patent citations	7.198	1	15.608	0	148
Patent baseline stock of patent citations from industry	5.787	0	13.239	0	137
Patent baseline stock of patent citations at origin	0.370	0	1.714	0	34
Patent baseline stock of patent citations at destination	0.231	0	1.966	0	53

Notes: The match is patent centric; that is, the control patent is always chosen from the same application year and the same issue year. In addition, control and treatment patents are matched on patent citation dynamics up to the year before the (possibly counterfactual) transition year. The cost of a very close, nonparametric match on patent characteristics is that author characteristics do not match closely. Imposing a close match on focal scientist age, gender, and overall productivity at baseline would result in a match rate which is unacceptably low.

scientists' location. As an example, only 6.82 percent of citations up to the baseline year have accrued at the origin location; the figure is 1.77 percent in the destination location.

Third, whereas citations at the origin location are well matched at baseline, this is not the case for citations at destination. In all cases, movers have accrued many more citations in the area they will soon transition to, relative to the citations that have accrued to the products of stayers in a location picked at random. This is consistent with the view that mobility events are jointly determined with expected spillovers of knowledge; for instance, because scientists who know your work are more likely to win the competition to lure you away. These baseline differences further justify our emphasis on identifying a closely matched set of control products.

Finally, the salience of industrial citers varies greatly across our measures of knowledge flows, accounting for only 3.61 percent of article-to-article

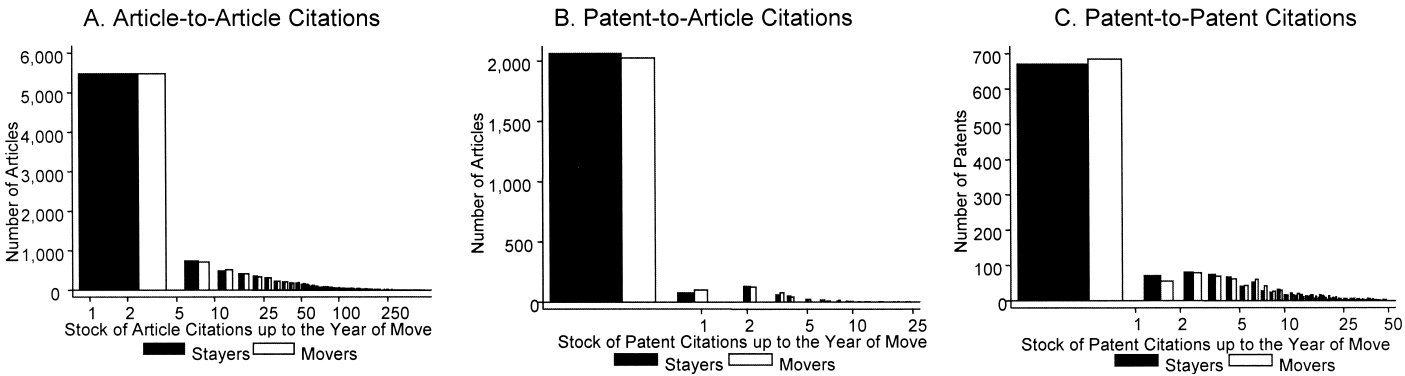

Fig. 2.6 Covariate balance at baseline

Note: We compute the cumulative number of citations for treatment and control articles/patents, respectively, up to the year that immediately precedes that of the professional transition for the superstar.

cites, but 80.72 percent of article-to-patent cites, and 83.04 percent of patent-to-patent citations.

2.2 Econometric Considerations

A natural starting point for a difference-in-difference (DD) analysis of the causal effect of labor mobility on knowledge flows is to conduct the statistical analysis using all product-year observations (treated and control) as the estimation sample. Since the mobility effect is mechanically correlated with the passage of time, as well as with an article's age, it is necessary to include life cycle and period effects, as is the norm in studies of scientific productivity (Levin and Stephan 1991).

In this framework, the control group that pins down the counterfactual vintage and calendar time effects for the products that were generated by scientists currently transitioning to new positions contains three categories of products: (a) those generated by movers who transitioned in earlier periods, (b) those generated by scientists who will move in the future, and (c) those generated by stayers. This approach is problematic insofar as products that appeared after a scientist has moved are not appropriate controls if the mobility event negatively affects the trend in their citations. If this is the case, fixed effects may underestimate the true effect of mobility.

To produce an analysis in which the control group consists solely of products associated with stayers, we instead perform the statistical analysis at the *product-pair* level. Specifically, the outcome variable is the *difference* between the citations received in a given year by a treated product and its associated control identified in the matching procedure previously described. Let i denote an article associated with a mover and let i' index the corresponding control product. Then our estimating equation relates $\Delta\text{CITES}_{ii't} = \text{CITES}_{it} - \text{CITES}_{i't}$ with the timing of mobility in the following way:

$$(1)\quad E[\Delta\text{CITES}_{ii't} \mid X_{ijt}] = \beta_0 + \beta_1\text{AFTER\_MOVE}_{jt} + f(\text{AGE}_{jt}) + \gamma_{ii'},$$

where AFTER_MOVE denotes an indicator variable that switches to one in the year focal scientist j moves, $f(\text{AGE})$ corresponds to a flexible function of the scientist's age, and the $\gamma_{ii'}$ correspond to product-pair fixed effects, consistent with our approach to analyze *changes* in the pair's citation rate following the move of investigator j.[8] We also run slight variations of this specification in which the dependent variable has been parsed so that we can break down citation flows by location or by citer type (i.e., industrial vs. academic citers).

There is another benefit to conducting the analysis at the product-pair level: since treated and control products always originate in the same year,

8. We do not need to include product vintage or year effects in the specification, since both products in the pair appeared in the same year, by construction.

experimental time and calendar time coincide, making it simple to display the results of the analysis graphically. The graphical approach is advantageous because it makes the essence of the empirical exercise transparent. The regression analysis, however, will prove useful when exploring interactions between the treatment effect and various star or product characteristics.

2.3 Results

2.3.1 Effect of Mobility on Citation Rates to Articles Published *After* the Move

As explained earlier, the bulk of our analysis focuses on citation flows to articles (respectively to patents) published (respectively issued) before the move so that we can separately identify the effect of mobility from that of correlated influences that might have an impact on the quality of the research itself. For example, mobility events such as those analyzed in this chapter might be driven by the availability of resources in the destination location, including laboratory equipment, trainees, or potential collaborators. From a descriptive standpoint, it is still interesting to examine the geographic spread of citations that accrue to products that postdate the mobility event, and these results are reported in figure 2.7. For the sake of brevity, we examine this for article-to-article citation flows only.[9]

We pair articles written by superstar movers with articles written by superstar stayers who are observationally quite similar at the time of the mobility event, so that the match is both "article-centric" and "scientist-centric." The scientist-level covariates used to create the match are (a) year of highest degree (coarsened in three-year intervals); (b) gender; (c) NIH funding status (funded vs. not funded at the time of the move); and (d) the total number of citations having accrued by 2006 to all premove publications. This ensures that the scientists being compared are not only demographically similar, but also of comparable renown at the time of the (possibly counterfactual) move. In addition, we match on article characteristics, including the journal, the length of the authorship roster, the focal author's position, and the publication year. Descriptive statistics for the resulting sample of $2 \times 26{,}254 = 52{,}508$ articles are displayed in table 2.6.

In the three panels of figure 2.7, we display the difference in average citation trends for the article pairs in the sample (the solid line), along with a 95th confidence interval (the dashed lines). Panel A focuses on differential citation patterns at the origin location. Relative to articles by stayers, it appears that postmove research is cited less in the area the moving scientist departed from; this citation discount is small (less than one citation per year

9. Our discussant Adam Jaffe uses the metaphor of carefully examining the dirty bath water before throwing it out to focus on the (hopefully clean) baby.

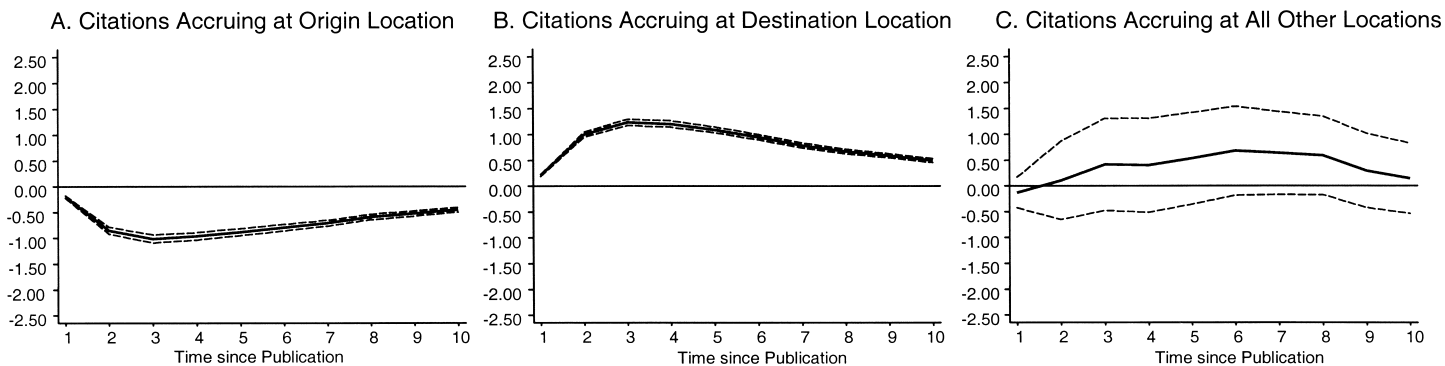

Fig. 2.7 Effect of professional transitions on article-to-article citation rates, by location (articles published *after* the move)

Note: Dynamics for the difference in yearly citations between movers' and stayers' matched articles written in the postmove period. Articles in each pair appeared in the same year and journal, and are also matched on focal scientist position on the authorship list, as well as overall number of authors. Further, the superstar authors are matched on gender, year of highest degree (in three-year bins), NIH funding at the time of the (possibly counterfactual) move, and cumulative number of citations for all articles published up to (and including) the year of the move.

In panel A, only citations accruing at the origin location are tallied. In panel B, only citations accruing at the destination location are tallied. For stayers, the destination location is chosen at random from among the set of locations that movers move to and that are separated by at least fifty miles from the staying star's actual location. In panel C, only citations accruing at all other locations are tallied.

Table 2.6 **Article-to-article citation flows: Descriptive statistics ($n = 2 \times 26{,}254$), articles published *after* the move**

	Mean	Median	Std. dev.	Min.	Max.
Journal articles by stayers					
Number of authors	4.915	4	4.077	1	255
Focal author is last	0.661	1	0.473	0	1
Article stock of citations up to 2006	273.818	133	542.514	1	22,336
Article publication year	1992.271	1992	6.208	1977	2003
Move year	1986.912	1987	6.212	1976	2002
Focal author graduation year	1970.270	1970	8.198	1931	1996
Focal author gender	0.023	0	0.149	0	1
Scientist citations at baseline	5,283	3,623	5,215	0	60,496
Scientist NIH funding at baseline	$3,828,267	$2,412,251	$5,297,164	$0	$101,678,352
Journal Articles by movers					
Number of authors	4.887	4	3.843	1	255
Focal author is last	0.661	1	0.473	0	1
Article stock of citations up to 2006	279.845	134	576.462	0	22,298
Article publication year	1992.271	1992	6.208	1977	2003
Move year	1986.912	1987	6.212	1976	2002
Focal author graduation year	1970.422	1970	7.990	1940	1996
Focal author gender	0.023	0	0.149	0	1
Scientist citations at baseline	5,248	3,584	5,157	0	51,174
Scientist NIH funding at baseline	$3,563,252	$2,306,315	$4,381,859	$0	$118,257,904

Notes: The match is both scientist centric and article centric. The control article is always chosen from the same journal in the same publication year. The control article is coarsely matched on the number of authors (exact match for one, two, and three authors; four or five authors; between six and nine authors; and more than nine authors). We also match on focal scientist's position in the authorship roster (first author; last author; middle author). In addition, the following individual covariates for the moving and staying stars match: gender, year of highest degree (in three-year bins), NIH funding status as of the moving year (funded vs. not); and total number of citations having accrued by 2006 to all premove publications (below the 10th percentile; between the 10th and 25th percentile; between the 25th percentile and the median; between the median and the 75th percentile; between the 75th and 95th percentile; between the 95th and 99th percentile; and above the 99th percentile).

on average), but it is enduring. Panel B repeats the same analysis for the destination location; we find the opposite pattern, in that postmove articles benefit from a lasting citation premium equal to less than one citation per year in the new location, relative to the number of citations accruing to the matched articles of stayers in a random location. Finally, Panel C examines citation outcomes in all other locations. Though the articles of movers appear to benefit from more "buzz" than those of stayers, this effect is both very small and imprecisely estimated.

From these results, it would appear that scientist mobility slightly shifts the allocation of citations across scientific areas without much of an impact on the diffusion process in the aggregate. Of course, because our controls for scientist-level and article-level quality are imperfect, we should resist the temptation to overinterpret these patterns. For instance, a citation discount at origin could mean that the superstar's former colleagues are quick to forget his or her research after the mobility event. But she or he may have

moved to a new location precisely because his or her research was delving into areas that appealed less to his or her old peers. In this case, the causality would flow from (expected) impact to job mobility, rather than in the direction we hypothesize. Similarly, at destination, the citation premium might reflect the interest of colleagues who extended an offer to the mover precisely in the expectation of deeper intellectual connections.

For these reasons, the rest of the chapter will focus on changes in citation rates following mobility events (and their allocation across geographic areas) for articles published before the move. This research design will enables us to better isolate the effect of mobility per se from that of correlated and competing influences.

2.3.2 Effect of Mobility on Citation Rates to Articles and Patents Published *Before* the Move

Our primary results are presented in figures 2.8 through 2.13. Table 2.7 presents estimates from simple ordinary least squares (OLS) regressions with article-pair fixed effects, corresponding to the earlier estimating equation. Robust standard errors, clustered at the scientist level, appear below the coefficient estimates in parentheses.

Article-to-Article Citation Flows. Panel A of figure 2.8 displays the citation dynamics corresponding to article-to-article flows, without disaggregating these flows by citer location or institutional type. It is clear from the picture that our matching procedure succeeded in identifying good control articles, since there is no evidence of deviation from zero in the years preceding the move. Moreover, there is a clear uptick in the rate of citations after the move, though it is modest in magnitude and relatively short-lived,

Table 2.7 **Effects of professional move on citation rates, by location**

	Article-to-article citations		Patent-to-article citations		Patent-to-patent citations	
	Origin (1a)	Destination (1b)	Origin (2a)	Destination (2b)	Origin (3a)	Destination (3b)
After appointment	0.026	0.069***	–0.026	0.007	–0.121***	0.041**
	(0.026)	(0.015)	(0.012)	(0.006)	(0.036)	(0.017)
Nb. of observations	175,715	175,715	41,114	41,114	21,221	21,221
Nb. of article pairs	10,249	10,249	2,435	2,435	1,417	1,417
Nb. of scientists	2,106	2,106	928	928	426	426
Adjusted R^2	0.295	0.323	0.157	0.125	0.215	0.214

Notes: Standard errors in parentheses, clustered by scientists. All specifications are estimated by OLS; the models include article-pair fixed effects.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

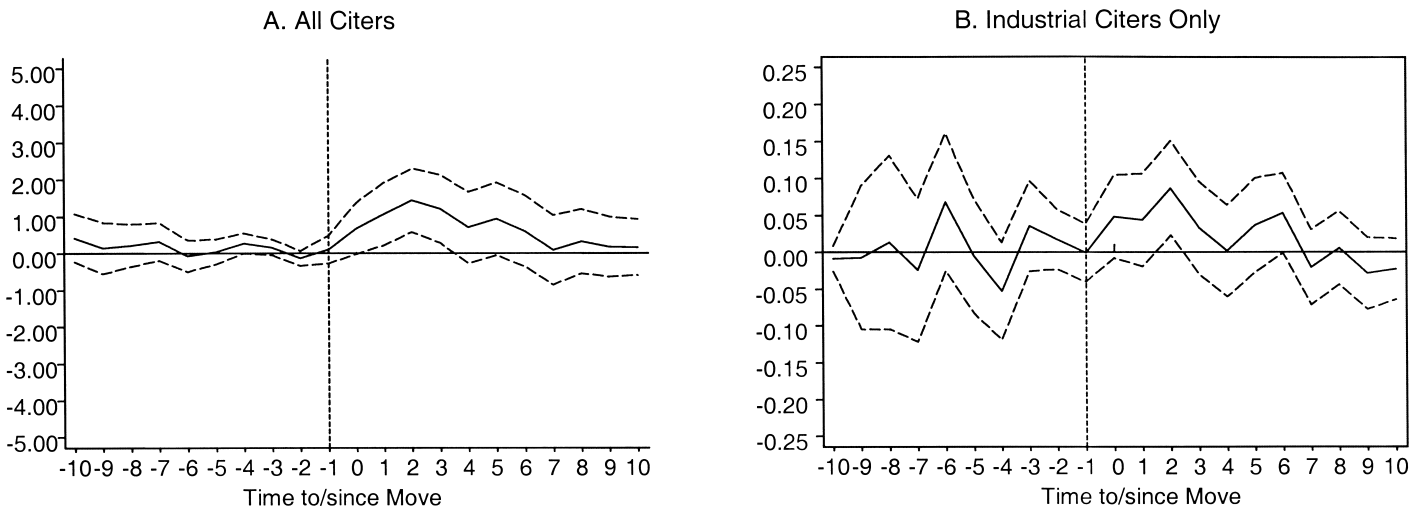

Fig. 2.8 Effect of professional transitions on article-to-article citation rates, by citing institution type (articles published *before* the move)

Notes: Dynamics for the difference in yearly citations between movers' and stayers' matched articles written in the premove period. Articles in each pair appeared in the same year and journal, and are also matched on focal scientist position on the authorship list, as well as overall number of authors. Further, control articles are selected such that the sum of squared differences in citations between control and treated article up to year $t_0 - 1$ is minimized—where t_0 is the year of (possibly counterfactual) move. In addition, when the year of publication is in the year prior to the move, the articles in each pair appeared not only in the same year, but also in the same month. Similarly, when the year of publication is in the penultimate year prior to the move, the articles in each pair appeared not only in the same year, but also in the same quarter.

fading out completely seven years later. Panel B examines whether the same patterns can be observed when restricting the outcome variable to article citations from industrial firms. The scale of the vertical axis is different, since these industrial cites account for a relatively tiny fraction of the total. Due to the paucity of the industrial citations, the results are very imprecise, though there is a very modest upward deviation from trend one year after the move.

Figure 2.9 display the results for citation flows disaggregated by citer location. Perhaps surprisingly, citations at the origin location do not appear to decline upon a star's departure (panel A). This lack of forgetting on the part of academics points to a capacity to absorb scientific knowledge that is disembodied from the producer of a particular idea. However, this view needs to be tempered in light of the results displayed in panel B, which focuses on citations accruing at the destination location. Relative to the flows in a random—but distant—location for the stayer, the level of flows is higher for movers at destination even before the move, with an upward trend starting two years before the move is effective. This provides strong evidence that academic superstars are, at least in part, "recruited for ideas" (Agarwal and Singh, forthcoming). Furthermore, this upward trend becomes more pronounced after the move, peaking two years later, but fading out only slowly over time. In other words, there is clear evidence that itinerant scientists circulate their old ideas in their new locations. The magnitude of this effect is not trivial: by the end of the observation period, movers have accrued more than twice as many citations to their old ideas at destination than stayers have in their counterfactual, random location.

Panel C examines citation dynamics in all locations, save for the origin and destination. One can discern a slight increase in citations after the move, though it is neither large nor precisely estimated. Yet, this should not be surprising if we think that mobility events give scientists looking for a new position an opportunity to give their ideas—old and new—a boost in exposure.

The asymmetry between the citation dynamics at location and origin strikes us as noteworthy, since it provides clear evidence that labor mobility increases the circulation of scientific ideas. If one espouses the view that knowledge flows are economically and socially valuable, then our results raise the intriguing possibility that scientists move too little, relative to what would lead to an optimal rate of scientific exploration. We return to this point in the discussion.

Tables 2.8 and 2.9 explore whether the magnitude of the treatment effect is affected by a number of article and scientist characteristics, at the origin and destination locations, respectively.[10] We do not discuss these in detail, since they tend to be quite noisy. Furthermore, with an unlimited number of

10. We do not repeat these analyses for patent-to-patent and patent-to-article citations, since they are sparse as is, and analyses that separate them into bins would be very noisy.

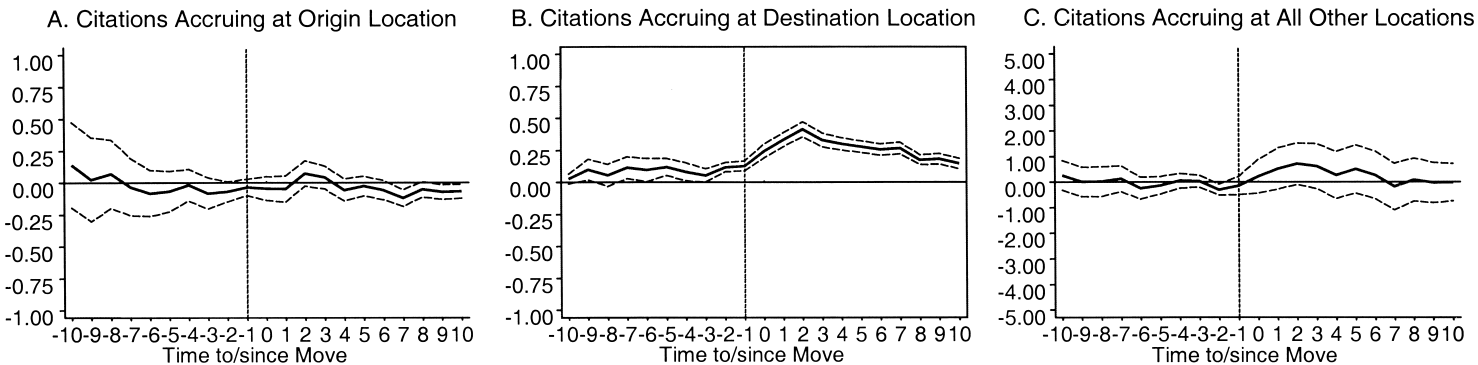

Fig. 2.9 Effect of professional transitions on article-to-article citation rates, by location (articles published *before* the move)

Note: Dynamics for the difference in yearly citations between movers' and stayers' matched articles written in the premove period. Articles in each pair appeared in the same year and journal, and are also matched on focal scientist position on the authorship list, as well as overall number of authors. Further, control articles are selected such that the sum of squared differences in citations between control and treated article up to year $t_0 - 1$ is minimized—where t_0 is the year of (possibly counterfactual) move. In addition, when the year of publication is in the year prior to the move, the articles in each pair appeared not only in the same year, but also in the same month. Similarly, when the year of publication is in the penultimate year prior to the move, the articles in each pair appeared not only in the same year, but also in the same quarter.

In panel A, only citations accruing at the origin location are tallied. In panel B, only citations accruing at the destination location are tallied. For stayers, the destination location is chosen at random from among the set of locations that movers move to and that are separated by at least fifty miles from the staying star's actual location. In panel C, only citations accruing at all other locations are tallied.

Table 2.8 **Effects of professional move on article-to-article citation rates at origin location**

	Novel vs. not		Young vs. old		Journal prestige	
	Novel (1a)	Not (1b)	Young (2a)	Old (2b)	Low JIF (3a)	High JIF (3b)
After appointment	–0.054	0.083**	–0.000	0.045	0.032	0.022
	(0.039)	(0.034)	(0.041)	(0.033)	(0.028)	(0.040)
Nb. of observations	80,165	95,550	82,580	93,135	84,114	91,601
Nb. of article pairs	3,713	6,536	4,524	5,725	4,884	5,365
Nb. of scientists	1,273	1,648	1,192	914	1,698	1,489
Adjusted R^2	0.243	0.320	0.290	0.299	0.269	0.305

	Pre- vs. post-Internet		Big vs. small status change		PI vs. non-PI pubs	
	1975–1994 (4a)	1995–2003 (4b)	Big (5a)	Small (5b)	First or last position (6a)	Middle position (6b)
After appointment	0.023	0.038	–0.014	0.037	0.045*	–0.026
	(0.028)	(0.067)	(0.057)	(0.029)	(0.025)	(0.066)
Nb. of observations	150,880	24,835	34,414	141,301	130,230	45,485
Nb. of article pairs	7,456	2,793	2,049	8,200	7,315	2,934
Nb. of scientists	1,782	564	417	1,689	1,872	1,200
Adjusted R^2	0.251	0.361	0.322	0.289	0.253	0.334

	Well-cited at baseline		Well-funded at baseline		Prolific patenter at baseline	
	No (7a)	Yes (7b)	No (8a)	Yes (8b)	No (9a)	Yes (9b)
After appointment	0.011	0.043	0.021	0.044	0.016	0.054
	(0.032)	(0.040)	(0.030)	(0.052)	(0.028)	(0.057)
Nb. of observations	88,823	86,892	131,548	44,167	135,818	39,897
Nb. of article pairs	5,240	5,009	7,787	2,462	7,477	2,772
Nb. of scientists	1,406	700	1,708	398	1,732	374
Adjusted R^2	0.276	0.306	0.285	0.322	0.275	0.328

Note: Standard errors in parentheses, clustered by scientists. All specifications are estimated by OLS; the models include article-pair fixed effects.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

potential contingencies, pure luck would dictate that at least some interaction effects would be statistically significant. In table 2.8, we find almost all interaction effects to be imprecisely estimated zeros.

New results, some reassuring, others more puzzling, emerge when focusing on citation flows at destination (table 2.9). We examine whether the mobility premium at destination varies with authorship credit for the focal

Table 2.9 **Effects of professional move on article-to-article citation rates at destination location**

	Novel vs. not		Young vs. old		Journal prestige	
	Novel (1a)	Not (1b)	Young (2a)	Old (2b)	Low JIF (3a)	High JIF (3b)
After appointment	0.036*	0.093***	0.035	0.094***	0.078***	0.063***
	(0.021)	(0.020)	(0.025)	(0.018)	(0.014)	(0.024)
Nb. of observations	80,165	95,550	82,580	93,135	84,114	91,601
Nb. of article pairs	3,713	6,536	4,524	5,725	4,884	5,365
Nb. of scientists	1,273	1,648	1,192	914	1,698	1,489
Adjusted R^2	0.237	0.360	0.343	0.305	0.256	0.345
	Pre- vs. post-Internet		Big vs. small status change		PI vs. non-PI pubs	
	1975–1994 (4a)	1995–2003 (4b)	Big (5a)	Small (5b)	First or last Position (6a)	Middle Position (6b)
After appointment	0.046***	0.164***	0.039	0.077***	0.077***	0.047
	(0.015)	(0.043)	(0.038)	(0.016)	(0.016)	(0.030)
Nb. of observations	150,880	24,835	34,414	141,301	130,230	45,485
Nb. of article pairs	7,456	2,793	2,049	8,200	7,315	2,934
Nb. of scientists	1,782	564	417	1,689	1,872	1,200
Adjusted R^2	0.295	0.361	0.355	0.315	0.313	0.342
	Well-cited at baseline		Well-funded at baseline		Prolific patenter at baseline	
	No (7a)	Yes (7b)	No (8a)	Yes (8b)	No (9a)	Yes (9b)
After appointment	0.077***	0.061**	0.078***	0.043**	0.066***	0.078**
	(0.016)	(0.025)	(0.018)	(0.026)	(0.016)	(0.034)
Nb. of observations	88,823	86,892	131,548	44,167	135,818	39,897
Nb. of article pairs	5,240	5,009	7,787	2,462	7,477	2,772
Nb. of scientists	1,406	700	1,708	398	1,732	374
Adjusted R^2	0.253	0.350	0.331	0.297	0.309	0.340

Notes: Standard errors in parentheses, clustered by scientists. All specifications are estimated by OLS; the models include article-pair fixed effects.

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

scientists. For this purpose, we exploit a robust social norm in the natural and physical sciences, whereby last authorship is systematically assigned to the principal investigator of a laboratory, first authorship is generally assigned to the junior author who was responsible for the actual conduct of the investigation (or, more rarely, to the principal investigator (PI) of a collaborating lab), and the remaining credit is apportioned to authors in the middle of

the authorship list, generally as a decreasing function of the distance from the extremities (Riesenberg and Lundberg 1990). We split the cited-article sample in two by consolidating the first and last authorship categories, and contrasting it with those article-pairs in which the focal scientists appear in the middle of the authorship list. We find clear evidence of a more pronounced mobility effect for article pairs in which the departing scientist is either first or last author. The evidence for middle-position authors is much smaller in magnitude. This is reassuring because the level of contribution of middle authors is often sufficiently small that one would not expect these old, marginal articles (from the point of view of the mover's overall corpus of work) to gain significant exposure at the new destination.

Second, we fail to detect a mobility premium of larger magnitude for the citations to the papers of superstars who shine particularly bright, regardless of the ways in which we seek to distinguish the elite from others who might be less accomplished (models 7a through 9b of table 2.9).

The anomalous result that bears mention pertains to the sample split corresponding to articles of recent versus older vintage. We separate the analysis for papers that appeared prior to and after 1995, a date that we pick as a marker for the Internet becoming ubiquitous in academia. We find that the mobility premium is four times higher for papers written in the Internet era than for papers published in the pre-Internet times. These results are inconsistent with the widespread belief that the diffusion of the Internet led to the "death of distance," though they should be interpreted cautiously since they may also reflect other changes over time.

Patent-to-Article Citation Flows. Figure 2.10 and 2.11 present the evidence on the second measure of knowledge flows, citations made to articles published in the open science literature in patents granted by the USPTO. In panel A of figure 2.10, we cannot detect any differential citation trend for the overall citations flowing to treated, rather than control articles. In fact, there is only the faintest evidence of a decline after the move. Panel B, which focuses on citations from industrial assignees alone, similarly shows no clear result.

The evidence on localization, presented in figure 2.11, is also relatively weak. This time, we observe a meaningful decline of citations at the origin location following the departure of a superstar, but this temporary dip is not pinned down precisely. Table 2.7, column (2a) presents the same analysis in regression form, but uses a longer postmove observation period, and constrains the mobility effect to be constant over time. In this case, we can detect a statistically significant decline equal to a quarter of a citation per year on average. Similarly, we observe a small increase in citations for treated articles at destination, relative to controls, but we cannot reject the hypothesis of a mobility premium equal to zero.

Patent-to-Patent Citation Flows. We employ the more traditional measure of knowledge flows—patent-to-patent citations—in the next batch of analyses, presented in figures 2.12 and 2.13. Once again, premove citation

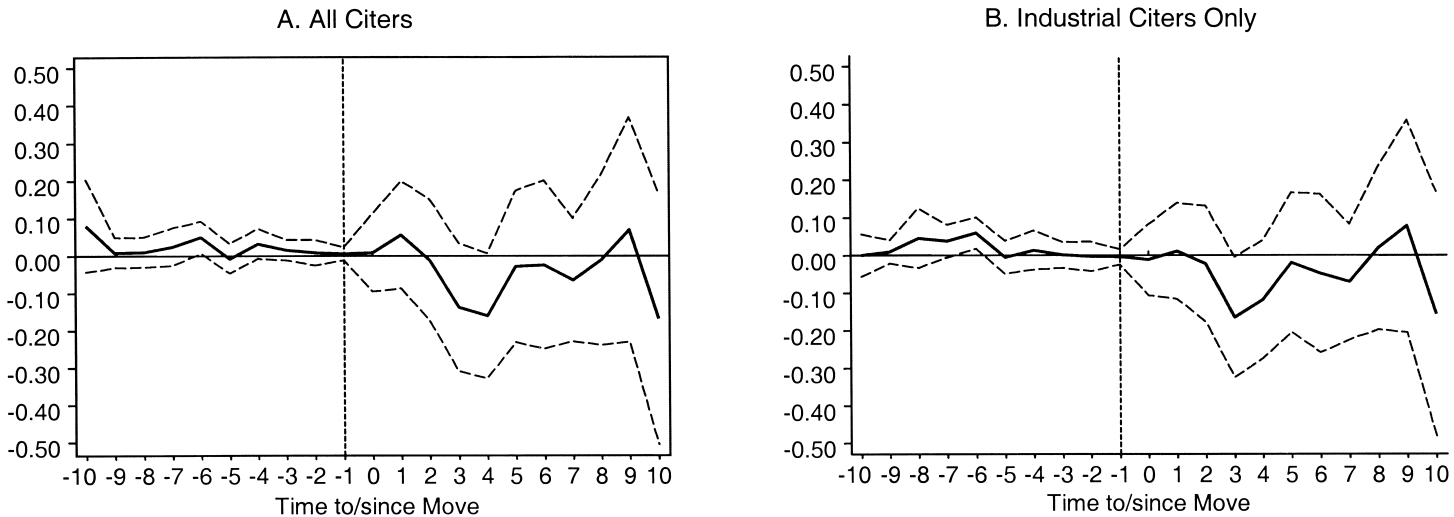

Fig. 2.10 Effect of professional transitions on patent-to-article citation rates, by citing institution type (articles published *before* the move)

Notes: Dynamics for the difference in yearly citations between movers' and stayers' matched articles written in the premove period. Articles in each pair appeared in the same year and journal, and are also matched on focal scientist position on the authorship list, as well as overall number of authors. Further, control articles are selected such that the sum of squared differences in citations between control and treated article up to year $t_0 - 1$ is minimized—where t_0 is the year of (possibly counterfactual) move. In addition, when the year of publication is in the year prior to the move, the articles in each pair appeared not only in the same year, but also in the same month. Similarly, when the year of publication is in the penultimate year prior to the move, the articles in each pair appeared not only in the same year, but also in the same quarter.

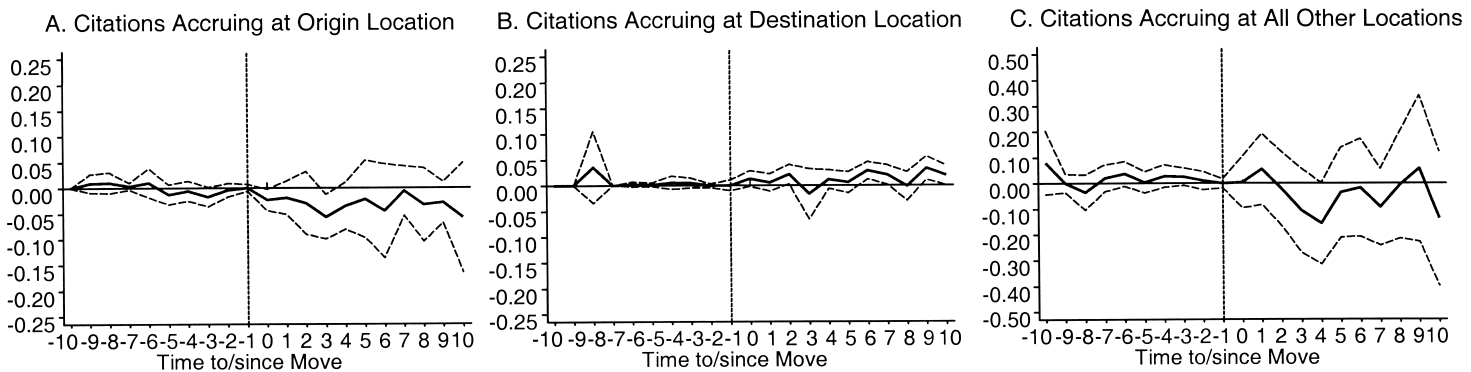

Fig. 2.11 Effect of professional transitions on patent-to-article citation rates, by location (articles published *before* the move)

Note: Dynamics for the difference in yearly citations between movers' and stayers' matched articles written in the premove period. Articles in each pair appeared in the same year and journal, and are also matched on focal scientist position on the authorship list, as well as overall number of authors. Further, control articles are selected such that the sum of squared differences in citations between control and treated article up to year $t_0 - 1$ is minimized—where t_0 is the year of (possibly counterfactual) move. In addition, when the year of publication is in the year prior to the move, the articles in each pair appeared not only in the same year, but also in the same month. Similarly, when the year of publication is in the penultimate year prior to the move, the articles in each pair appeared not only in the same year, but also in the same quarter.

In panel A, only citations accruing at the origin location are tallied. In panel B, only citations accruing at the destination location are tallied. For stayers, the destination location is chosen at random from among the set of locations that movers move to and that are separated by at least fifty miles from the staying star's actual location. In panel C, only citations accruing at all other locations are tallied.

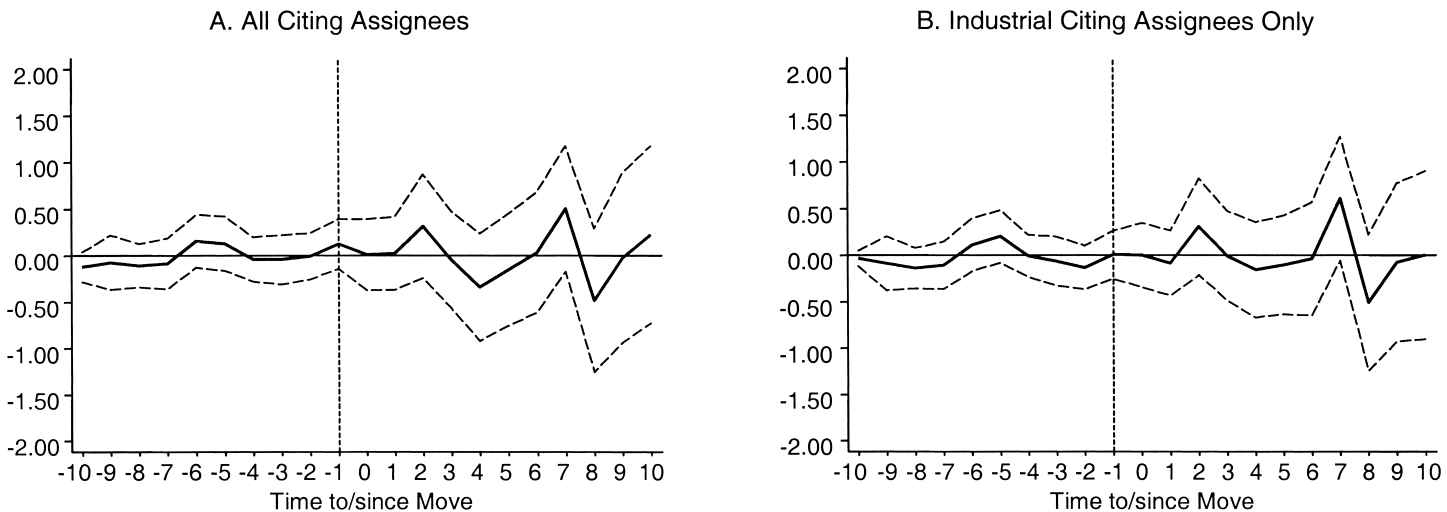

Fig. 2.12 Effect of professional transitions on patent-to-patent citation rates, by citing assignee type (patents issued *before* the move)

Notes: Dynamics for the difference in yearly citations between movers' and stayers' matched patents issued in the premove period. Patents in each pair share the same application and issue years. Further, control patents are selected such that the sum of squared differences in citations between control and treated patents up to year $t_0 - 1$ is minimized—where t_0 is the year of (possibly counterfactual) move.

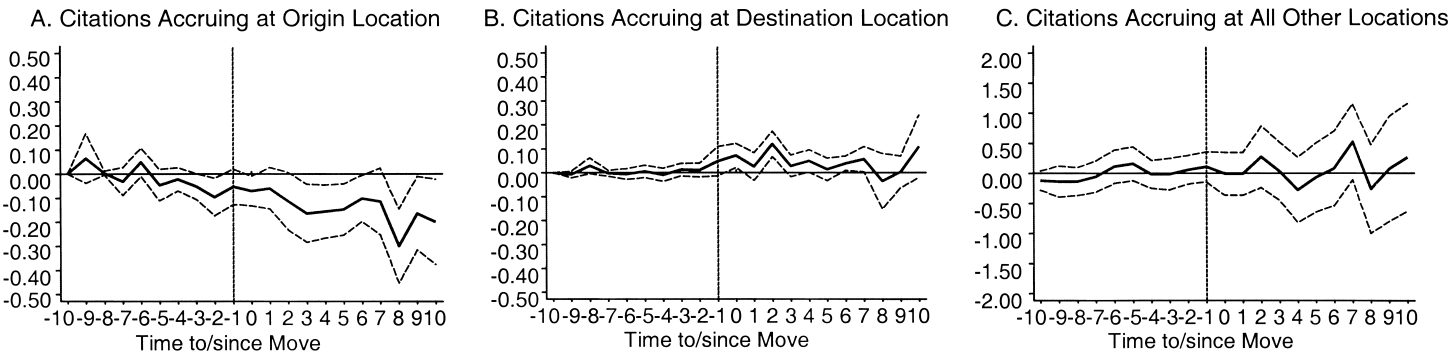

Fig. 2.13 Effect of professional transitions on patent-to-patent citation rates, by location (patents issued *before* the move)

Note: Dynamics for the difference in yearly citations between movers' and stayers' matched patents patents in the premove period. Patents in each pair share the same application and issue years. Further, control patents are selected such that the sum of squared differences in citations between control and treated patents up to year $t_0 - 1$ is minimized—where t_0 is the year of (possibly counterfactual) move.

In panel A, only citations accruing at the origin location are tallied. In panel B, only citations accruing at the destination location are tallied. For stayers, the destination location is chosen at random from among the set of locations that movers move to and that are separated by at least fifty miles from the staying star's actual location. In panel C, only citations accruing at all other locations are tallied.

dynamics appear very similar up until the year of the move, which is expected given the extensive efforts we deployed in our search for appropriate control patents. There is no evidence that mobility increases citation flows overall (figure 2.12).

The localization effects presented in figure 2.13 are much more dramatic. First, there is a decline in the rate of citations in the location of origin, which becomes more pronounced over time and shows no sign of abating even ten years after the scientist has departed (panel A). This may reflect the importance of physical proximity to university laboratories for helping industrial firms develop the inventions of academic entrepreneurs (Zucker, Darby, and Brewer 1998; Audretsch and Stephan 1996). However, this interpretation is undermined by evidence that the onset of this decline precedes the move by almost four years (though this preexisting downward trend is small in magnitude and imprecisely estimated).

The upward trend at destination (panel B) is not quite as dramatic, but clear. Here again, there is some weak evidence of anticipation, with citations being slightly higher for movers in the baseline year at destination, relative to stayers. It is therefore difficult to distinguish between the view that physical proximity to academic entrepreneurs begets absorptive capacity, from the alternative perspective that a scientist's assessment that the local industrial base has grown stale (or at least less receptive to his/her ideas) triggers mobility.

2.4 Discussion

In this chapter, we examine the impact of geography on knowledge transfer by exploiting professional transitions within the academic life sciences coupled with publication and patent citation data over time. The results reveal a rather nuanced story. Consistent with models of localized knowledge diffusion, we find strong evidence that publication-to-publication citations (to papers published before the move) rise at destination locations after the move takes place. We also find, however, persuasive evidence of a legacy effect at the origin institution—citation rates do not decline after the scientist departs. While the findings on the patent side are less conclusive than those on publications alone, they reveal a slightly different role for geography. Here again citations at the destination location rise (or at least remain the same) after the move, while citations at the origin location appear to fall, particularly for patent-to-patent citations.

The normative implications of our findings are not straightforward, especially since there may be first-order effects of job mobility we do not observe. Nonetheless, we offer some broader speculations here. Let's begin with a deeper look at the publication-to-publication citation results. A surge in citations at the new location with little drop off at the old location underscores the importance of scientist interactions, but also makes clear that these interactions are not easily forgotten. Since the sharing and recombin-

ing of existing ideas is viewed as an essential component in the innovation process (Weitzman 1998; Burt 2004; Simonton 2004), might our evidence suggest that scientists are moving too little?

The answer to this will, of course, depend upon the degree to which scientists internalize the impacts of their location decisions, but suboptimal levels of mobility seem likely. Nearly all of the costs of moving are borne privately, yet much of the credit associated with new scientific discoveries is apportioned out narrowly to the lead scientists on that project, leaving researchers with incentives that appear too weak from a societal perspective. While the best way to address this limited mobility is unclear, it has potentially large and important implications for the rate, and especially the direction of technological innovation within the economy, and eventually for economic growth.

The analysis of patent citations—most of which are generated by biomedical firms—suggests a distinct knowledge production process within industry. The output of local talent is most influential when it remains local. That ideas are quickly forgotten after a scientist departs suggests an important role for face-to-face interactions. One possible explanation for this finding is that the limited absorptive capacity within most firms necessitates a substantive dialogue with academic scientists in order to translate scientific output into something more useful for organizations concerned with its translation into marketable products. Such dialogues are clearly less costly with local talent, especially if the fruitful search for ideas is not one that is narrowly circumscribed around a well-defined issue. The opportunities that are lost when a scientist departs, however, are not entirely clear. Even if firms are abandoning science that the academy believes is still useful, what is the proper benchmark here? The academy and industry may simply value different types of ideas. Even still, some ideas that firms should value are likely to fall off the radar screen when scientists depart, offering at least some temperance to the idea that the innovative costs of scientist mobility are negligible.

These conjectures assume the construct validity of our measures: that publication-publication citations actually measure knowledge flows among academics, and that patent-patent and patent-publication citations actually measure academic industry spillovers. In the spirit of recognizing measurement difficulties (see e.g., Kuznets and Schmookler in the 1962 volume) we acknowledge these are assumptions. For example, numerous scholars of bibliometrics have noted the ceremonial function of publication citations (Merton 1968; MacRoberts and MacRoberts 1996). While we have interpreted the finding that there is little forgetting of superstars research after a move as evidence that face-to-face interaction may not be so necessary for knowledge flows, it could instead reflect that the scientists continue to be cited for ceremonial reasons. A related explanation: if citations are less about intellectual influence than just knowing about research (MacRoberts

and MacRoberts 1996), we may not expect any decay after professional transitions.

Similar concerns could be raised about citations in patents. As Jaffe and Trajtenberg and others have emphasized, the claim that these citations reflect real knowledge flows, or spillovers, is only an assumption. Survey work (Jaffe, Trajtenberg, and Fogarty 2002) suggests they are noisy measures. Recent analyses (Alcácer, Gittelman, and Sampat 2009; Sampat 2010; Hegde and Sampat 2009) on the importance of patent examiners in generating these citations may undermine the notion that they are true knowledge flows. One of the reasons for using patent-to-publication citations is that these are less affected by examiner influence (Lemley and Sampat 2010) and potentially better measures of knowledge flows (Roach and Cohen 2010), though here too there are questions of whether applicants have incentives to disclose all relevant knowledge, and only relevant knowledge (Cotropia, Lemley, and Sampat 2010). All this granted, it is difficult to construct an explanation of our "forgetting" result for patent-to-patent and patent-to-publication citations that is driven only by incentives to cite (or citation practices).

Appendix A

Criteria for Delineating the Set of 10,450 "Superstars"

We present additional details regarding the criteria used to construct the sample of 10,450 superstars.

Highly Funded Scientists. Our first data source is the Consolidated Grant/Applicant File (CGAF) from the US National Institutes of Health (NIH). This data set records information about grants awarded to extramural researchers funded by the NIH since 1938. Using the CGAF and focusing only on direct costs associated with research grants, we compute individual cumulative totals for the decades 1977 to 1986, 1987 to 1996, and 1997 to 2006, deflating the earlier years by the Biomedical Research Producer Price Index.[11] We also recompute these totals excluding large center grants that usually fund groups of investigators (M01 and P01 grants). Scientists whose totals lie in the top ventile (i.e., above the 95th percentile) of either distribution constitute our first group of superstars. In this group, the least well-funded investigator garnered $10.5 million in career NIH funding, and the most well-funded $462.6 million.[12]

11. http://officeofbudget.od.nih.gov/UI/GDPFromGenBudget.htm.

12. We perform a similar exercise for scientists employed by the intramural campus of the NIH. These scientists are not eligible for extramural funding, but the NIH keeps records of

Highly Cited Scientists. Despite the preeminent role of the NIH in the funding of public biomedical research, the previous indicator of superstardom biases the sample toward 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 broad scientific field.[13]

Top Patenters. We add to these groups academic life scientists who belong in the top percentile of the patent distribution among academics—those who were granted 17 patents or more between 1976 and 2004.

Members of the National Academy of Sciences. We add to these groups academic life scientists who were elected to the National Academy of Science between 1975 and 2007.

MERIT Awardees of the NIH. Initiated in the mid-1980s, the MERIT Award program extends funding for up to five years (but typically three years) to a select number of NIH-funded investigators "who have demonstrated superior competence, outstanding productivity during their previous research endeavors and are leaders in their field with paradigm-shifting ideas." The specific details governing selection vary across the component institutes of the NIH, but the essential feature of the program is that only researchers holding an R01 grant in its second or later cycle are eligible. Further, the application must be scored in the top percentile in a given funding cycle.

Former and Current Howard Hughes Medical Investigators. Every three years, the Howard Hughes Medical Institute selects a small cohort of mid-career biomedical scientists with the potential to revolutionize their respective subfields. Once selected, HHMIs continue to be based at their institutions, typically leading a research group of ten to twenty-five students, postdoctoral associates, and technicians. Their appointment is reviewed every five years, based solely on their most important contributions during the cycle.[14]

Early Career Prize Winners. We also included winners of the Pew, Searle, Beckman, Rita Allen, and Packard scholarships for the years 1981 through 2000. Every year, these charitable foundations provide seed funding to between twenty and forty young academic life scientists. These scholarships

---

the number of internal projects each intramural scientist leads. We include in the elite sample the top ventile of intramural scientists according to this metric.

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

14. See Azoulay, Graff Zivin, and Manso (2011) for more details and an evaluation of this program.

are the most prestigious accolades that young researchers can receive in the first two years of their careers as independent investigators.

Appendix B

Linking Scientists with Their Journal Articles

The source of our publication data is PubMed, a bibliographic database maintained by the US National Library of Medicine that is searchable on the web at no cost.[15] PubMed contains over 14 million citations from 4,800 journals published in the United States and more than 70 other countries from 1950 to the present. The subject scope of this database is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering that inform research in health-related fields. In order to effectively mine this publicly available data source, we designed Publication Harvester, an open-source software tool that automates the process of gathering publication information for individual life scientists (see Azoulay, Stellman, and Graff Zivin 2006 for a complete description of the software). Publication Harvester is fast, simple to use, and reliable. Its output consists of a series of reports that can be easily imported by statistical software packages.

This software tool does not obviate the two challenges faced by empirical researchers when attempting to link accurately individual scientists with their published output. The first relates to what one might term "Type I error," whereby we mistakenly attribute to a scientist a journal article actually authored by a namesake; the second relates to "Type II error," whereby we conservatively exclude from a scientist's publication roster legitimate articles.

Namesakes and Popular Names. PubMed does not assign unique identifiers to the authors of the publications they index. They identify authors simply by their last name, up to two initials, and an optional suffix. This makes it difficult to unambiguously assign publication output to individual scientists, especially when their last name is relatively common.

Inconsistent Publication Names. The opposite danger, that of recording too few publications, also looms large, since scientists are often inconsistent in the choice of names they choose to publish under. By far the most common source of error is the haphazard use of a middle initial. Other errors stem from inconsistent use of suffixes (Jr., Sr., 2nd, etc.), or from multiple patronyms due to changes in spousal status.

To deal with these serious measurement problems, we opted for a labor-

15. http://www.pubmed.gov/.

intensive approach: the design of individual search queries that relies on relevant scientific keywords, the names of frequent collaborators, journal names, as well as institutional affiliations. We are aided in the time-consuming process of query design by the availability of a reliable archival data source, namely, these scientists' curriculum vitaes (CVs) and biosketches. PublicationHarvester provides the option to use such custom queries in lieu of a completely generic query (e.g., "azoulay p" [au] or "sampat bn" [au]). As an example, one can examine the publications of Scott A. Waldman, an eminent pharmacologist located in Philadelphia, PA, at Thomas Jefferson University. Waldman is a relatively frequent name in the United States (with 208 researchers with an identical patronym in the Association of American Medical Colleges (AAMC) faculty roster); the combination "waldman s" is common to three researchers in the same database. A simple search query for "waldman sa" [au] OR "waldman s" [au] returns 302 publications at the time of this writing. However, a more refined query, based on Professor Waldman's biosketch returns only 210 publications.[16]

The previous example also makes clear how we deal with the issue of inconsistent publication names. PublicationHarvester gives the end-user the option to choose up to four PubMed-formatted names under which publications can be found for a given researcher. For example, Louis J. Tobian, Jr. publishes under "tobian 1," "tobian 1 jr," and "tobian 1j," and all three names need to be provided as inputs to generate a complete publication listing. Furthermore, even though Tobian is a relatively rare name, the search query needs to be modified to account for these name variations, as in ("tobian l" [au] OR "tobian 1j" [au])

We are confident that this labor-intensive customization ensures the accuracy of our superstar scientists' bibliomes.

Appendix C

Linking Scientists with Their Patents

A number of recent efforts have been devoted to assigning unique identifiers to inventors in the US Patent Data (Trajtenberg, Shiff, and Mclamed 2006; Marx, Strumsky, and Fleming 2009). Rather than relying on recursive algorithms that help group together patents issued to the same inventors, we make use of the richness of our data to improve the quality of the matched inventor/invention links.

In a first step, we eliminate from the set of potential patents all patents

16. (((“waldman sa” [au] NOT (ether OR anesthesia)) OR (“waldman s” [au] AND (murad OR philadelphia [ad] OR west point [ad] OR wong p [au] OR lasseter kc [au] OR colorectal))) AND 1980:2010 [dp])

issued in classes that appear unrelated to the life sciences, writ large. Second, we focus on the set of superstars with relatively rare names, and automate the match with the patent data by declaring as valid any link in which (a) the inventor's full name matches, and (b) at least one patent assignee matches with one of the scientist's employer, past or present. We then relax these constraints one at a time, examining potential matches by hand. Using knowledge about the research of these scientists stemming from their biographical records, we then pass judgement on the validity of these more uncertain matches. The same procedure is repeated for the set of inventors with common names, though these records often require the inspection of each potential patent to ascertain whether they correspond to legitimate or spurious matches.

Following Thursby, Fuller, and Thursby (2009), we find that many patents associated with the elite scientists in our sample are not assigned to their employer, but rather unassigned, or assigned solely to an industrial firm. As a result, we are very careful to inspect manually records for which the inventor name matches that of one of our superstars, but there is no assignee information to match with the available biographical record for this individual.

One objection to this linking procedure is that it is ad hoc, and difficult to replicate across different empirical contexts. Moreover, it is very labor intensive, and therefore would not scale up to a much larger sample of inventors. Yet, we suspect that using prior knowledge about the direction of an inventor's research to link them precisely with their patented inventions results in higher-quality matches.

Appendix D

Linking PubMed References to USPTO Patents

Determining whether patents cite publications is more difficult than tracing patent citations: while the cited patents are unique seven-digit numbers, cited publications are free-form text (Callaert et al. 2006). Moreover, the USPTO does not require that applicants submit references to literature in a standard format. For example, Harold Varmus's 1988 *Science* article "Retroviruses" is cited in twenty-nine distinct patents, but in numerous different formats, including Varmus; "Retroviruses" *Science* 240:1427–1435 (1988) (in patent 6794141) and Varmus et al., 1988, *Science* 240:1427–1439 (in patent 6805882). As this example illustrates, there can be errors in author lists and page numbers. Even more problematic, in some cases certain fields (e.g., author name) are included, in others they are not. Journal names may be abbreviated in some patents, but not in others.

To address these difficulties, we developed a matching algorithm that compared each of several PubMed fields—first author, page numbers, volume, and the beginning of the title, publication year, or journal name—to all references in all biomedical and chemical patents issued by the USPTO since 1976. Biomedical patents are identified by technology class, using the patent class-field concordance developed by the National Bureau of Economic Research (Jaffe and Trajtenberg 2005). We considered a dyad to be a match if four of the fields from PubMed were listed in a USPTO reference.

Overall, the algorithm returned 558,982 distinct PMIDs (unique article identifiers in PubMed) cited in distinct 172,815 patents. Since it necessarily relied on probabilistic rather than exact matches, we also tested it across a sample of references where we were confident the match to the PubMed data was accurate. Specifically, we sampled 200 references from the biomedical/chemical patents, and two research assistants and one of the authors (Sampat) manually investigated whether the references had associated PMIDs. Sampat carefully reviewed and adjudicated any cases where there was disagreement among the three coders.

Manual matching, while cumbersome, provides an extremely reliable match, a gold standard against which we can gauge the algorithm. The algorithm returned the correct PMID information for 86 percent of the references. There were no false positives: if our manual match returned no PMID, neither did our algorithm. And in almost all cases, if the algorithm generated a PMID, it was the correct one. But for 14 percent of the references there were false negatives; that is, a PMID was found via the manual match, but none was found via the algorithm. While these errors are unlikely to be related to any variables of interest, we can also test robustness of any results obtained using these data using matches from a more liberal implementation of the algorithm (based on matching three rather than four elements of the PubMed record to the patent references), which returns fewer false negatives but more false positives.

Choosing between the loose and strict algorithms involves making tradeoffs between the Type I and Type II errors. In the analyses following, we rely primarily on the strict algorithm, erring on the side of understating the extent to which patents cite the biomedical literature.

Appendix E

Construction of the Product Control Group

We detail the "coarse exact matching" (CEM) procedure implemented to identify the sample of control products from among the universe of products

associated with stayers. As opposed to methods that rely on the estimation of a propensity score, CEM is a nonparametric procedure.[17]

In its basic outline, the matching procedure is very similar across the three measures of knowledge flows; whether we are focused on journal articles or patents, the sample of control products is constructed such that the following two conditions are met:

1. Treated articles/patents exhibit no differential citation trends relative to control products up to the time of mobility.
2. Treated and control articles/patents match on a number of time-invariant article characteristics.

However, implementation details vary with cited and citing product type, as explained later.

Journal Articles. We identify controls based on the following set of covariates: (1) year of publication; (2) specific journal (e.g., *Cell* or the *New England Journal of Medicine*); (3) number of authors (the distribution is coarsened into six bins: one, two, three, four or five, between six and nine, and ten or more authors); (4) focal-scientist position on the authorship list (first author, middle author, or last author). In the case of articles published in the year immediately preceding appointment, the list of matching covariates is expanded to also include the month of publication. In the case of articles published two years before appointment, the list of matching covariates is expanded to also include the quarter of publication. To ensure that premove citation trends are similar, we proceed in two steps. First, we also match on cumulative number of citations at baseline, coarsened into 7 strata (0 to 10th; 10th to 25th; 25th to 50th; 50th to 75th; 75th to 95th; 95th to 99th; and above the 99th percentile). However, we have found that this is not enough to eliminate premove citation trends. As a result, we select all control articles that match according to the previous covariates, and pick among those potential matches a single article that further minimizes the sum of squared differences in the number of citations up until the year before the year of move.

Patents. We identify controls based on the following set of time-invariant covariates: (1) year of issue; (2) year of application; and (3) main patent class. To ensure that premove citation trends are similar, we match on cumulative number of citations at baseline, coarsened into 4 strata (0 to 50th; 50th to 95th; 95th to 99th; and above the 99th percentile).

17. A propensity score approach would entail estimating the probability that the scientists in the data move in a given year, and then using the inverse of this estimated probability to weight the data in a second stage analysis of the effect of mobility on subsequent citation rates. However, because citations occur at the article level, achieving covariate balance by weighting the data by the scientist-level likelihood of moving, even if the determinants of mobility were observable, would not resolve the problem of controlling for article-level quality.

Coarse Exact Matching. We create a large number of strata to cover the entire support of the joint distribution of the covariates mentioned earlier. Each observation is allocated to a unique strata. We then drop from the data all observations corresponding to strata in which there is no treated article and all observations corresponding to strata in which there are less than five potential controls.

The procedure is coarse because we do not attempt to precisely match on covariate values; rather, we coarsen the support of the joint distribution of the covariates into a finite number of strata, and we match a treated observation if and only if a control observation can be recruited from this strata. An important advantage of CEM is that the analyst can guarantee the degree of covariate balance ex ante, but this comes at a cost: the more fine-grained the partition of the support for the joint distribution (i.e., the higher the number of strata), the larger the number of unmatched treated observations.

We implement the CEM procedure year by year, without replacement. Specifically, in move year t, $1976 \leq t \leq 2004$, we do the following:

1. Eliminate from the set of potential controls all products published by stayers who have collaborated with movers prior to year t
2. For each year of publication/issue $t - k$, $1 \leq k \leq 10$
    a. Create the strata
    b. Identify within strata a control for each treated unit; break ties at random
    c. Repeat these steps for year of publication/issue $t - (k + 1)$
3. Repeat these steps for year of appointment $t + 1$

Sensitivity Analyses. The analyst's judgement matters for the outcome of the CEM procedure insofar as he must draw a list of reasonable covariates to match on, as well as decide on the degree of coarsening to impose. Therefore, it is reasonable to ask whether seemingly small changes in the details have consequences for how one should interpret our results.

Nonparametric matching procedures such as CEM are prone to a version of the "curse of dimensionality" whereby the proportion of matched units decreases rapidly with the number of strata. For instance, requiring scientist-level characteristics to match in addition to article-level characteristics would result in a match rate below 10 percent, which seems to us unacceptably low.

However, we have verified that slight variations in the details of the implementation (e.g., varying slightly the number of cutoff points for the stock of citations) have little impact on the basic results we present. To conclude, we feel that CEM enables us to identify a population of control products appropriate to guard against the specific threats to identification mentioned in section 2.1.4.

References

Aghion, Philippe, Mathias Dewatripont, Fiona Murray, Julian Kolev, and Scott Stern. 2009. "Of Mice and Academics: Examining the Effect of Openness on Innovation." NBER Working Paper no. 14819. Cambridge, MA: National Bureau of Economic Research, March.

Aghion, Philippe, and Peter Howitt. 1992. "A Model of Growth through Creative Destruction." *Econometrica* 60 (2): 323–51.

Agrawal, Ajay, Iain Cockburn, and John McHale. 2006. "Gone But Not Forgotten: Labor Flows, Knowledge Spillovers and Enduring Social Capital." *Journal of Economic Geography* 6 (5): 571–91.

Agrawal, Ajay, and Rebecca Henderson. 2002. "Putting Patents in Context: Exploring Knowledge Transfer from MIT." *Management Science* 48 (1): 44–60.

Agrawal, Ajay, and Jasjit Singh. 2011. "Recruiting for Ideas: How Firms Exploit the Prior Inventions of New Hires." *Management Science* 57 (1): 129–50.

Alcácer, Juan, and Michelle Gittelman. 2006. "How Do I Know What You Know? Patent Examiners and the Generation of Patent Citations." *Review of Economics and Statistics* 88 (4): 774–79.

Alcácer, Juan, Michelle Gittelman, and Bhaven Sampat. 2009. "Applicant and Examiner Citations in U.S. Patents: An Overview and Analysis." *Research Policy* 38 (2): 415–27.

Almeida, Paul, and Bruce Kogut. 1999. "Localization of Knowledge and the Mobility of Engineers in Regional Networks." *Management Science* 45 (7): 905–17.

Audretsch, David B., and Paula E. Stephan. 1996. "Company-Scientist Locational Links: The Case of Biotechnology." *American Economic Review* 86 (3): 641–52.

Azoulay, Pierre, Waverly Ding, and Toby Stuart. 2009. "The Effect of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output." *Journal of Industrial Economics* 57 (4): 637–76.

Azoulay, Pierre, Joshua Graff Zivin, and Gustavo Manso. 2011. "Incentives and Creativity: Evidence from the Academic Life Sciences." *RAND Journal of Economics* 42 (3): 527–54.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang. 2010. "Superstar Extinction." *Quarterly Journal of Economics* 125 (2): 549–89.

Azoulay, Pierre, Andrew Stellman, and Joshua Graff Zivin. 2006. "PublicationHarvester: An Open-Source Software Tool for Science Policy Research." *Research Policy* 35 (7): 970–4.

Azoulay, Pierre, Toby Stuart, and Yanbo Wang. 2011. "Matthew: Effect or Fable?" Massachusetts Institute of Technology, Working Paper.

Belenzon, Sharon, and Mark Schankerman. 2010. "Spreading the Word: Geography, Policy and University Knowledge Diffusion." Center for Economic and Policy Research. Discussion Paper no. 8002.

Blackwell, Matthew, Stefano Iacus, Gary King, and Giuseppe Porro. 2009. "CEM: Coarsened Exact Matching in Stata." *The Stata Journal* 9 (4): 524–46.

Branstetter, Lee. 2005. "Exploring the Link Between Academic Science and Industrial Innovation." *Annales d'Economie et de Statistique* 79–80 (Suppl): 119–42.

Burt, Ronald S. 2004. "Structural Holes and Good Ideas." *American Journal of Sociology* 110 (2): 349–99.

Callaert, Julie, Bart Van Looy, Arnold Verbeek, Koenraad Debackere, and Bart Thijs. 2006. "Traces of Prior Art: An Analysis of Non-Patent References Found in Patent Documents." *Scientometrics* 69 (1): 3–20.

Cech, Thomas R. 2005. "Fostering Innovation and Discovery in Biomedical Research." *Journal of the American Medical Association* 294 (11): 1390–3.

Cockburn, Iain M., and Rebecca M. Henderson. 1998. "Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery." *Journal of Industrial Economics* 46 (2): 157–82.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh. 2002. "Links and Impacts: The Influence of Public Research on Industrial R&D." *Management Science* 48 (1): 1–23.

Cole, Jonathan R., and Stephen Cole. 1972. "The Ortega Hypothesis." *Science* 178 (4059): 368–75.

Cotropia, Christopher A., Mark Lemley, and Bhaven Sampat. 2010. "Do Applicant Citations Matter? Implications for the Presumption of Validity." Columbia University. Working Paper.

Dasgupta, Partha, and Paul David. 1994. "Towards a New Economics of Science." *Research Policy* 23 (5): 487–521.

de Solla Price, Derek J. 1963. *Little Science, Big Science.* New York: Columbia University Press.

Fallick, Bruce, Charles A. Fleischmann, and James B. Rebitzer. 2006. "Job Hopping in Silicon Valley: Some Evidence Concerning the Micro-foundations of a High Technology Cluster." *Review of Economics and Statistics* 88 (3): 472–81.

Furman, Jeffrey, and Scott Stern. 2011. "Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Knowledge Production." *American Economic Review* 101 (5): 1933–63.

Hegde, Deepak, and Bhaven Sampat. 2009. "Applicant Citations, Examiner Citations, and the Private Value of Patents." *Economics Letters* 5 (3): 287–9.

Henderson, Rebecca, Adam Jaffe, and Manuel Trajtenberg. 2005. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment." *American Economic Review* 95 (1): 461–4.

Henderson, Rebecca, Luigi Orsenigo, and Gary P. Pisano. 1999. "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.

Jaffe, Adam B., and Manuel Trajtenberg. 1999. "International Knowledge Flows: Evidence from Patent Citations." *Economics of Innovation and New Technology* 8 (1): 105–36.

Jaffe, Adam B., and Manuel Trajtenberg. 2005. *Patents, Citations, and Innovations.* Cambridge, MA: MIT Press.

Jaffe, Adam B., Manuel Trajtenberg, and Michael S. Fogarty. 2002. "The Meaning of Patent Citations: Report on the NBER/Case-Western Reserve Survey of Patentees." In *Patents, Citations, and Innovations,* edited by Adam B. Jaffe and Manuel Trajtenberg, 379–401. Cambridge, MA: MIT Press.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson. 1993. "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations." *Quarterly Journal of Economics* 108 (3): 577–98.

Krugman, Paul. 1991. "Increasing Returns and Economic Geography." *Journal of Political Economy* 99 (3): 483–99.

Kuznets, Simon. 1962. "Inventive Activity: Problems of Definition and Measurement." In *The Rate and Direction of Economic Activity: Economic and Social Factors,* Universities-National Bureau Committee for Economic Research and the Committee on Economic Growth of the Social Science Research Councils, 19–52. Princeton, NJ: Princeton University Press.

Lemley, Mark, and Bhaven Sampat. 2010. "Examiner Experience and Patent Office Outcomes." Columbia University. Working Paper.

Levin, Sharon G., and Paula E. Stephan. 1991. "Research Productivity Over the Life Cycle: Evidence for Academic Scientists." *American Economic Review* 81 (1): 114–32.

Lotka, Alfred J. 1926. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* 16 (12): 317–23.

MacRoberts, M. H., and Barbara MacRoberts. 1996. "Problems of Citation Analysis." *Scientometrics* 36 (3): 435–44.

Marburger, John H. 2005. "Wanted: Better Benchmarks." *Science* 308 (5725): 1087.

Marx, Matthew, Debbie Strumsky, and Lee Fleming. 2009. "Mobility, Skills, and the Michigan Non-compete Experiment." *Management Science* 55 (6): 875–99.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159 (3810): 56–63.

Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigation.* Chicago: University of Chicago Press.

Riesenberg, Don, and George D. Lundberg. 1990. "The Order of Authorship: Who's on First?" *Journal of the American Medical Association* 264 (14): 1857.

Roach, Michael, and Wesley M. Cohen. 2010. "Patent Citations as Measures of Knowledge Flows from Public Research." University of North Carolina. Working Paper.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5): S71–S102.

Sampat, Bhaven. 2010. "When Do Applicants Search for Prior Art?" *Journal of Law and Economics* 53 (2): 399–416.

Simonton, Dean Keith. 2004. *Creativity in Science: Chance, Logic, Genius, and Zeitgeist.* New York: Cambridge University Press.

Stephan, Paula E. 2010. "The Economics of Science." In *Handbook of The Economics of Innovation,* edited by Bronwyn H. Hall and Nathan Rosenberg, 217–73. Amsterdam: North-Holland.

Thompson, Peter, and Melanie Fox-Kean. 2005. "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment." *American Economic Review* 95 (1): 450–60.

Thursby, Jerry, Anne W. Fuller, and Marie Thursby. 2009. "US Faculty Patenting: Inside and Outside the University." *Research Policy* 38 (1): 14–25.

Trajtenberg, Manuel, Gil Shiff, and Ran Melamed. 2006. "The 'Names Game': Harnessing Inventors' Patent Data for Economic Research." NBER Working Paper no. 12479. Cambridge, MA: National Bureau of Economic Research, September.

Weitzman, Martin L. 1998. "Recombinant Growth." *Quarterly Journal of Economics* 113 (2): 331–60.

Zucker, Lynne G., Michael R. Darby, and Jeff Armstrong. 1999. "Intellectual Capital and the Firm: The Technology of Geographically Localized Knowledge Spillovers." NBER Working Paper no. 4946. Cambridge, MA: National Bureau of Economic Research, April.

Zucker, Lynne G., Michael R. Darby, and Marilynn B. Brewer. 1998. "Intellectual Human Capital and the Birth of U.S. Biotechnology Enterprises." *American Economic Review* 88 (1): 290–306.