Evolving Networks of Scientists and Engineers: Strategic Complementarity and Government Policy**

By

James D. Adams, Rensselaer Polytechnic Institute and NBER
Ameya M. Haté, Akamai Technologies
J. Roger Clemmons, United Healthcare

February 2025: Revision 1

Comments Welcome

Keywords: Social and Economic Networks, Science and Engineering, Research & Development, Defense Procurement, Government Policy

JEL Codes: D24, I23, J24, O33

Corresponding Author: James D. Adams, Department of Economics, Rensselaer Polytechnic Institute, 3504 Russell Sage Laboratory, Troy, NY 12180-3590. Telephone: 1-518-396-7163, E-mail: adamsj14@rpi.edu.

** This research began during a visit to the National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, Massachusetts, USA. We thank the NBER for its support. Thomson-Reuters IP & Science, now Clarivate Analytics, advised us concerning Web of Science™, which forms the core of our data. Malik Magdon-Ismail, Professor of Computer Science at Rensselaer Polytechnic Institute, provided crucial advice as regards the algorithms essential for identifying scientists in this paper. We are grateful to Julie Hotchkiss, Daniel Klinenberg, and Peter Orazem for valuable criticisms of earlier drafts. NSF award SMA-1064227 helped fund the project.

Abstract

Employing Web of Science™ we construct citation networks for U.S. research institutions during 1981-2011. Machine learning methods identify scientists, institutional affiliations, and numbers of scientists by nodes comprised of institutions and fields. Linked by citation edges, the nodes form networks of scientific research.

Edges define centralities that measure importance of nodes. We exploit the centralities to study the demand for scientists and the supply of research. With randomly assigned centralities as instruments, we find that centralities increase scientific employment and research output, so research in a node is a strategic complement to research elsewhere. Thus, government policy creates a system of positive externalities.

Supplementary variables describe the defense R&D cycle and their effect on employment in defense institutions. The interactions are often positive and statistically significant for employment of long-term researchers in contractor firms. During the contraction phase, long-term research scientists decline due to separations from employers. During the expansion phase, employment recovers.

I. Introduction

Since World War II the United States has made large public investments in scientific research in the pursuit of domestic prosperity and national security[1]. And indeed, these investments have paid off in terms of U.S. innovation in agriculture and forestry, computers and software, semiconductors, defense and aerospace, and more recently biotechnology, medicine, and information technology[2]. These beneficial outcomes have stimulated U.S. economic growth to the present day.

This paper explores a mechanism that sustains these outcomes. Government support of science has created a network of research institutions that include universities, firms, federal agencies, federal laboratories, and non-profit biomedical research centers. Distinguished by field and location, the institutions form nodes of networks that are connected by citations and collaborations, with citations the primary focus of this paper[3]. Also, the nodes are connected by migration as scientists leave one node to join another, which we incorporate in the research labor forces of nodes of the networks.

The goal of this paper is to understand contributions of nodes of the networks to each other. Does a node's importance, defined by its centrality, contribute to its scientific employment and research output and that of other nodes? If the answer is yes, research in one node increases research in another and research is a strategic complement (Bulow, Geanakoplos, and Klemperer, 1985). In this case science policy creates a thriving system of network externalities. If the answer is no, research in one node diminishes research elsewhere, research is a strategic substitute, and the externalities are negative. The overwhelming evidence in Section VI is that scientific research in one node does increase research in

[1] For agriculture, state and federal support of research began with the Morrill Act of 1862, which established the system of state agricultural experiment stations.

[2] Mowery, Nelson, Sampat, and Ziedonis (2004), Chapter 2, provide an historical overview of state and federal support of research at U.S. institutions. See the references cited there. Government support of research began in the nineteenth century, though federal support accelerated during and after World War II (Adams, 2010). Langlois and Mowery (1996) discuss federal government support of software. While semiconductors were invented in industry, federal support led to extensive semiconductor research in universities, especially the University of California (Lécuyer, 2014). Gross and Sampat (2023) trace long run effects of the World War II-Era Office of Scientific Research and Development on technology clusters throughout the U.S. For the rise in the university share in federal R&D during the postwar period, see Adams (2010) and National Science Board (various years).

[3] We emphasize citations over collaborations because citations are highly correlated with collaborations at the level of nodes. This makes it hard to identify a separate effect of collaborations at this level. We report collaboration centrality effects in Appendix D, where their contribution is found to be negligible.

other nodes and does form a strategic complement to that research. Therefore, science policy contributes to knowledge externalities among nodes of the network by supporting research in each node. Since these are external effects there is a significant role for public policy towards science beyond funding of individual institutions and fields because the allocation of federal funds influences observed externalities.

Before we can study the networks we must build them from scratch using the Web of Science ™ database of scientific papers. We arrive at the networks by a series of steps, which Sections II and III explain. The networks contain 255 top U.S. research institutions[4]. Up to twenty fields of science occur in an institution-location. Together institution, location, and field define a node, so that an institution can have multiple nodes distributed by location and field. Finally, the data span the period 1981-2011. The network is linked together by citations that crisscross among the nodes[5].

To make the process of data construction clear from the beginning, we provide an overview in this introduction. Section II.A defines a list of alternative names for each U.S. institution. This step is easier than the rest because Thomson-Reuters provided us with unifying name files for nonprofit organizations. However, these unification files are not available for firms. In this case we begin by identifying the top 100 R&D firms as defined by Compustat at the start of the project. We construct our own unification files using institution names in Web of Science, which we match by name with Top 100 R&D firm names in Compustat and with subsidiary names of firms in Mergent Online. The result is a set of items with aliases of Top 100 R&D firms as defined as Web of Science institutions.

Section II.B conducts the second data identification step. It identifies individual scientists and distinguishes them from author names. This is the most involved step because it relies on iterative scoring methods. The work begins by assigning author names sharing the same last name and first initial, known as a name-hash, to a bucket. The total number of buckets for U.S. institutions is 1,732,238. This is less

[4] The 255 institutions are comprised of 148 universities, 59 firms, 25 federal laboratories, nine federal agencies, and 14 nonprofit biomedical research centers. The 107 non-university institutions have 194 locations, so there are a total of 342 (148+194) institution-locations in the networks.

[5] The number of citation links, which are pairs of citing and cited institutions, locations, and fields, is 55,007,212. These are not the number of citations, but the number of links carrying the citations. The number of collaboration links is 1,570,262. This again is the number of links, not the number of collaborations.

than the number of scientists since multiple scientists share the same bucket. From a computational perspective a bucket can be thought of as a community of non-overlapping identities sharing the same name-hash, who are in fact different scientists. To split the community into distinct identities we apply a widely used algorithm in computer science. This is the Fast Modularity Community Structure Inference Algorithm due to Clauset, Newman, and Moore (2004).

Before we can apply the algorithm we must cluster author names on different papers under an ID number that provisionally represents a scientist. For this purpose, a similarity score is needed. The similarity score for a pair of author names is based on seven separate scores for author, address, item, journal issue, coauthors, citations, and two interactions, one for same institution and same subject, the other for same subject and same coauthor. A lower cutoff is needed to split the graph for each bucket into separate groups. The Fast Modularity algorithm sets the lower cutoff.

We evaluated the scoring mechanism twice before applying it to the complete data set. First, we evaluated the scoring mechanism on a small sample to identify shortcomings. We revised the initial scoring mechanism based on the findings. Second, we used a hand-collected sample of sixty eminent scientists from **American Men and Women of Science** whose more than five thousand papers were known in advance. We compared predicted to actual authorship to assess the revised scoring mechanism and revise it a second time. We found errors of a few percentage points that were comparable to prior research on patents. Details are in Section II.B and Appendix A. Only then did we apply the scoring mechanism and Fast Modularity Community algorithm to the entire dataset.

Section II.C is the final data identification step. It assigns scientists identified in the author identification step to a sequence of institutional affiliations. The sequence defines the employment and research history of the scientist. We use Dijkstra's Algorithm to construct the sequence. The idea of Dijkstra's Algorithm is this. Assuming that moving between jobs is costly, Dijkstra's Algorithm minimizes the number of employment changes. We evaluate the algorithm by comparing actual changes to predicted changes using the sample of sixty eminent scientists described above. We describe the errors in Appendix B. Most scientists rarely change affiliations and errors in changes in employment changes

are small. More important, errors in the timing of employment change, in which researchers move too soon or too late, are independent. The errors cancel out when we aggregate scientists by node and year, as we do in the empirical work.

Once these steps are taken, we find that there are 5,850,091 scientists in U.S. institutions. This is 3.4 scientists per bucket[6]. There are 6,350,323 scientist-affiliations, or 1.1 affiliations per scientist. This concludes the discussion of the data identification steps in Section II.

Section III builds on Section II. It constructs counts of researchers, papers, and citations received by node and year. Recall that nodes are institutions, locations, and fields[7]. Combined with citations between nodes, this section defines contents of the nodes forming citation networks.

We conduct a few more refinements. For tractability we break the networks up by six-year intervals from 1981-1986, 1986-1991, …, 2006-2011. And to avoid small nodes where scientific research is inherited from graduate work, we require that a university-field node have a graduate program and that it publishes at least sixty papers during each six-year interval[8]. For non-university nodes, during a six-year interval we require at least four hundred publications in an institution and field, and at least one hundred papers in an institution-field-location.

Next, we calculate two centrality measures. These are citation in-degree centrality and citation PageRank centrality[9]. There are advantages to the in-degree centrality. First, the citation networks are weighted networks, where more important nodes are cited more and receive more weight. The in-degree centrality retains significant differences between fields and institutions in the number of incoming citations. Second, network theorists prefer in-degree centrality to out-degree centrality because it indicates

[6] Since there are 5,850, 091 scientists at the conclusion, and since there are 1,732,238 buckets, the ratio of the two is about 3.4 scientists per bucket.
[7] Fractional papers written at a particular institution, location, field and year are the share of its authors on each paper summed over all papers involving the node and year. The definition is the same for future citations received. Choice of fractional papers over whole papers avoids double counting of papers and citations.
[8] We experimented with a lower bound of thirty papers on university departments with results like those reported in Section VI. Including smaller departments makes trivial difference to the estimates.
[9] We compute other measures, but we focus on in-degree and PageRank centralities because of their attractive properties.

influence over neighboring nodes. And as Section III will point out, in-degree is related to network density.

The PageRank centrality also has several advantages. For one, it is readily computed on citation networks. For another, it avoids misattribution of centrality due to a node being cited by other central nodes since citations are divided by out-degree. And because citation networks resemble networks of web pages, for which PageRank was developed, application of the PageRank centrality is appropriate. The Bonacich-Katz centrality is not divided by out-degree. It therefore misattributes centrality due to a node being cited by other central nodes. This is a serious drawback.

Compared to in-degree, PageRank has the advantage of including indirect citations as well as direct citations by neighbors. However, PageRank is normalized so that it sums to 1.0 across nodes. This is a disadvantage because normalization flattens field effects compared to the in-degree centrality.

Section IV presents network level data. Because density and average in-degree centrality are proportional, and density rises over time in all fields, average in-degree centrality rises as well. Besides this, we find that the number of network components has declined along with network fragmentation. These observations suggest that science networks have become more efficient in transmitting information.

Section V presents an economic analysis of networks. Part A models the behavior of individual scientists. It assumes that scientist allocate time and effort between administration and research to maximize income. To explain behavior of nodes of the network Part B introduces a version of the linear model of social interactions. It incorporates income-maximizing actions of scientists from Part A while defining the objective of the node and deriving the demand for research scientists. This framework yields linear best-response demand functions for research scientists, where centrality increases the demand for scientists if research in different institutions are strategic complements. In this model the reduced form of research output depends on employment of research scientists, contingent on the best response.

Building on Section V, Section VI estimates best-response demand functions for research scientists and research output equations. The estimates use panel data drawn from the networks. The estimates employ 2SLS, and 3SLS methods, where the equation errors are robust or else clustered by nodes.

Crucially, this section develops a method for random assignment of centralities from other nodes in a past network. For example, Harvard Medical School, the largest university research organization in medicine, could be assigned a centrality from any other node in university medicine in a past network. Random assignment breaks the correlation between the degree centrality and omitted variables that bias the centrality coefficient in OLS regressions. The 2SLS estimates show a positive and highly significant effect of the network centrality on research scientists free of simultaneity bias, confirming strategic complementarity of research in the network. In two equation models that jointly estimate best response demand functions for research scientists and the research output equation, research is again a strategic complement across the network. And in the research output equation, employment of research scientists, instrumented by federal policy towards funding of different sciences, significantly increases research output.

The paper builds on the knowledge production function literature pioneered by Griliches (1979, 1992) and the citation literature in Jaffe and Trajtenberg (2002), which deal with the production and exchange of information. A related literature of general production functions includes Olley and Pakes (1996), Levinsohn and Petrin (2003), and Ackerberg, Caves, and Frazer (2015). This paper is in the tradition of knowledge production functions, but it differs in that it concerns *network* knowledge production functions. Accordingly, the instruments rely on random assignment from other nodes in prior networks and on federal policy towards different sciences in the networks. While there have been studies of industry R&D networks by Goyal and Moraga-González (2001), König, Liu and Zenou (2019), and others to our knowledge there have been no studies of scientific networks that encompass all institutions and identify research scientists at each node of the network as this study does.

In the study of networks, one of the first considerations is measurement of centralities of nodes of the networks. Centralities assess importance of nodes and rewards from participating in a network[10]. They are

[10] For examples of centralities see Katz (1953), Freeman (1977), Bonacich (1987), Brin and Page (1998), Borgatti (2005), Newman (2005), Borgatti and Everett (2006), Brandes (2008), and "Page Rank" (**Wikipedia**, accessed November 17, 2019). The key point about centralities is that they measure distinct aspects of importance in networks. And since networks vary in their properties, not all centralities can be computed for a given network.

part of a larger agenda of network measurement that is discussed in Wasserman and Faust (1994), Jackson (2008) and Newman (2010).

Network economics concerns optimizing behavior in networks, including the extent to which agents rely on their own effort versus efforts of neighbors in networks. Manski (1993) characterizes two channels by which networks transmit their effects, distinguishing between exogenous contextual effects, which do not depend on a neighboring agent's effort, and endogenous effects that do. This lays the groundwork for studying relationships of strategic complementarity and substitution between agents. Jackson and Wolinsky (1996) study deliberate formation of network links in a utility-maximizing framework, and the various network forms that emerge from the costs and benefits of forming links.

Ballester, Calvó-Armengol, and Zenou (2006) introduce the linear model of social interactions, in which agents choose effort to maximize a quadratic objective, and neighbors' efforts are assumed to be strategic complements or substitutes. Under strategic complements, an agent's effort can be shown to be proportional to their importance as captured by the Bonacich (1987) centrality. Calvó-Armengol, Patacchini, and Zenou (2009) apply the model to education, while Ballester, Calvó-Armengol, and Zenou (2010) apply the model to delinquency from school. Bramoullé, Djebbari, and Fortin (2009) develop instrumental variable methods for the model using networks where the error term is idiosyncratic and specific to an agent, rather than affecting all agents. This allows information on characteristics of neighbors of local neighbors to become instruments for characteristics of neighbors.

Bramoullé, Kranton, and D'Amours (2014) extend the model to strategic substitutes, specifically public goods, where provision of the goods by neighbors, while it benefits the agent, encourages free riding. A more extreme case is R&D rivalry, in which more effort by neighbors harms the agent and could lower incentives to do R&D. A related application by Immorlica, Kranton, Manea, and Stoddard (2017) is a modern interpretation of Veblen's relative consumption hypothesis. Agents take costly actions with direct benefits that confer indirect benefits of social status. Neighbors' prosperity is costly to agents because it lowers social status. Acemoglu, García-Jimeno, Robinson (2015) apply the linear model to the role of state capacity in the economic development of Colombia. This is a network model since state

capacities of neighboring municipalities are strategic complements to local state capacity. Blume, Brock, Durlauf, and Joyaraman (2015) develop a Bayes-Nash equilibrium framework for the linear model under uncertainty, in which agents have partial knowledge of neighbors. König, Liu, and Zenou (2019) use the linear model to study R&D alliances in industry. Other network models use alternative specifications of social interactions[11]. For references to recent literature, see Jackson, Rogers, and Zenou (2017)[12].

Last, the linear framework has been extended to nonlinear social interactions by Belhaj, Bramoullé, and Deroïan (2014) and Boucher, Rendall, Ushchev, and Zenou (2024). Boucher et al. (2024) replace the Linear-in-Means model of peers (Manski, 1993) with a CES function of neighbors' individual attributes. For example, instead of mean Grade Point Average they estimate a CES function of the vector of neighbor's Grade Point Averages. They find that the CES model rejects the Linear-in-Means model for most attributes. For our citation (information) networks, the in-degree centrality is heavily weighted towards influential nodes, so it departs from the Linear-in-Means model. It accepts imperfect substitutability between research in different nodes and assigns greater weight to influential research.

The rest of the paper proceeds as follows. Section II conducts data identification steps needed for citation networks composed of scientific institutions, locations, and field nodes. Section III builds the networks of citations at the node level and divides the data into separate networks lasting six years over 1981-2011. Section IV presents descriptive findings at the network level. Section V introduces models of the research scientist and the nodes where scientists are employed, and derives best response demand functions for research scientists. Section VI estimates demand functions for research scientists and supply equations for research output. Section VII summarizes, concludes, and suggests further directions for research.

[11] See Battaglini and Patacchini (2018) for the allocation of pressure group contributions, which increases with the Katz-Bonacich centrality of legislators. See Fainmesser and Galeotti (2020) for the equilibrium subsidization of influencers of other consumers in a social network of consumers.

[12] In recent years, the literature has branched off to consider production networks and the fragility of networks. For production networks, see Carvalho and Tahbaz-Salehi (2019). For a review of network fragility, see Eliott and Golub (2022). These literatures are distinct from the line of research pursued by this paper.

II. Data Identification and Construction

This section identifies institutions, scientists, and institutional affiliations of scientists in The Web of Science™ database[13]. This can be determined using machine learning, because each paper is associated with author names, addresses, institutional affiliations, journal names, journal classifications by science field, and publication dates.

Three sources of uncertainty prevent a solution and must be addressed. First, variation in organization names conceals the identities of institutions. Second, author names do not distinguish scientists from other scientists with a similar name. Third, authors are not matched to institutions except for single institution papers. The organization identification, name disambiguation, and affiliation algorithms described in this section solve these problems up to a margin of error common to all such efforts. With these approximate solutions, the location of scientists by institution and field is solved, opening the data to further analysis.

Two more steps define the networks. These are construction of nodes of the networks and construction of edges between nodes in the networks. Section III undertakes these steps.

A. Assignment of Research Organizations to Institutions

The Web of Science™ data contain twenty-eight million items for the period 1981-2011 from research organizations around the world. The scope of this project is bounded by U.S. research institutions. After these are defined, the items, which are mostly scientific papers, are limited to a subset associated with U.S. research organizations.

Organization names occur as variants consisting of abbreviations and alternate names. Fortunately, Thomson-Reuters provided us with unification files that contain name variants and distinguishing attributes like location that we can use to assign items to parent U.S. nonprofit institutions.

[13] Thomson Reuters Intellectual Property & Science, a unit of Thomson Reuters, produced Web of Science™. In October 2016, Intellectual Property & Science was sold to Onex Corporation and Baring Private Equity Asia. The new company was renamed Clarivate Analytics. In this paper we address the company using the name of the original parent, Thomson-Reuters, though the firm and the division have since parted ways.

Thomson-Reuters provided us with 407 unification files that represent more institutions than we ultimately included in the study, because some of the institutions are small and publish too intermittently to be retained for analysis. The files contain six columns with an extra seventh column that is occasionally populated. The first two contain organization name and parent organization name. The next three columns specify city, state, and country. The sixth column contains name variants for an organization. The seventh column contains entries that are to be used with component names when names are ambiguous. Using these data, we assigned organization names to parent nonprofit research institutions.

To accomplish the same goal for parent firms, we extracted names of the Top 100 R&D firms in the 2012 edition of Compustat at the outset of the project, and we collected names of subsidiaries of these parent firms that were listed under Mergent Online™. Using subsidiary names, we constructed organizational directories for firms that we matched by name to institutions in The Web of Science™.

B. Disambiguation of Author Names

The next step is to identify which author's names, given similar names on different papers, stand for the same person. This establishes the identity of the authors who are scientists in this study. This process is known as entity resolution or disambiguation[14]. Authors sharing the same last name and first initial—called a name hash—are treated as potentially the same author, even though the name hash in fact represents multiple authors. The set of authors with the same name hash forms a bucket. In this study the total number of buckets is 1,732,238, and the largest bucket contains 24,332 entries. To uncover the identities of persons contained in a bucket, we develop a scoring mechanism based on seven scores that together raise the probability of a true match on the same last name and first initial in different publications. Since there is no formal theory behind scoring, our algorithm is developed using training

[14] The data are too voluminous to disambiguate using manual curation. However, our repeated hand checks of samples of persons and papers whose identity was known beforehand found that our machine learning approach was accurate within a few percentage points.

data sets for which identities are known in advance. Total scores are defined for pairs of publications based on the following seven scores.

1. Author Score

Author score uses first name, second initial, suffix, and author key words for matching. If first names are not present in two publications, a score of zero is awarded. If full first names are present and match, five points are scored, while ten points are penalized if first names do not match. The same scoring mechanism is applied to the second initial and suffix. For keywords, the score is one point for each matching keyword. The final author score is the sum of all points, with the maximum possible being ten. All higher scores are truncated to ten points.

2. Address Score

Address score is based on email address and physical address. The score is zero if either email is missing from a pair of documents. If email addresses match, we assign ten points since this match indicates identity. If both emails exist but do not match, we penalize the score five points. The penalty is set low because email addresses do change.

We treat physical addresses differently, depending on whether the address is in the U.S. or not. If both physical addresses are in the U.S., then we extract city, state and postal code. A match on the state field results in two points, a match on the city field results in three points, while a match on the postal code field results in a score of five points. However, if a pair of addresses do not belong to the U.S., then we break the physical address into words. Each word overlap for a candidate pair yields one point. While doing this we have to account for word overlaps that do not imply similarity and occur frequently. For this purpose, we discard words such as ‘sci,’ ‘USA,’ ‘japan,’ ‘hosp,’ ‘england,’ and so on. We also discard articles and stop words. These words are typically dropped when checking similarity between documents. The final address score is the sum of these scores truncated to twenty points.

3. Item Score

Item score uses item title, language, item keywords and organization fields. For matching purposes, the procedure breaks up title and organization into words. Each match is worth one point, and high frequency and stop words are discarded. A score of one point is awarded if both documents are in English, while three points are scored if the two items are in the same non-English language, since other languages occur less frequently and are better indicators of similarity. A penalty of three points is imposed if languages do not match. One point is scored for every item keyword match. If both publications have been assigned to the same U.S. institution, five points are scored. The final item score is the sum of all individual scores truncated to fifteen.

4. Issue Score

Issue score uses issue title and issue subject area. An issue's title is scored in the same way as the item's title, as already described, while subject area is scored in the same way as item keyword. The sum of these scores is truncated to five.

5. Coauthor Score

If two author names have coauthors with the same last name and first initial, that is, the same name hash, then this is considered a match. One coauthor match is scored as five points whereas a match of if two coauthors is assigned 8 points. A coauthor match of size more than two results in ten points.

6. Citation Score

Within a bucket last name and first initial must match so two items share the same name hash. Under these conditions, if one item cites another item, then the algorithm assigns a score of 10 points between the two items, since this is likely to be a self-citation.

7. Interaction Score

If two items share the same institution and subject area, five points are scored, while another five points are scored if they share the same subject area, and they share a coauthor. These interaction scores depend on simultaneous matches. For example, the probability is exceptionally low that there would be two individuals with the same last name and first initial working in the same subject area at a given institution. The interaction scores give more weight to pairs of documents that exhibit such a property.

The final similarity score between the two publications is the sum of the seven scores, for which the maximum is set to 80. Notice that the first six scores use independent criteria, so a high combined score is a strong indication that two papers sharing the same name hash are written by the same person. The interaction score tunes the fit by taking simultaneous matches into account.

Since there is a score for each pair of publications, the set of linked publications forms a complete graph. We define a cutoff value to determine the subset of true matches among the set of same last names and first initials forming the bucket. We set the value of the cutoff from the performance of the Fast Modularity clustering algorithm (Clauset, Newman, and Moore, 2004), which breaks the graph into separate groups, each group representing the same person. By performance, we mean that hand-checking samples where identities of authors are known in advance finds that a given cutoff has the lowest percentage of errors. After repeated testing, we find that a cutoff on the similarity score of at least 20 out of 80 points fit the data best and identify any two publications as being by the same person.

Appendix A reports findings using our algorithm from an analysis of over 5,000 hand-collected papers by sixty distinguished scientists, whose careers are documented in **American Men and Women of Science** (31st edition, Gale Cengage, 2013)[15]. We find a lumping error of 6.6%, where separate scientists are falsely treated as the same person. We find a splitting error of 2.8%, where scientists are falsely treated as separate but are the same person. For the UC Berkeley-Fung Institute patent database, Li (2013) finds a lumping error of 4.6% and a splitting error of 2.0%. Since patent data include first names, while

[15] Access to **American Men and Women of Science** was provided by Cabot Science Library, Harvard University.

scientific papers mostly do not, our algorithm can be said to perform comparably to the Fung Institute algorithm.

Once author disambiguation has been achieved, we use Thomson-Reuters' journal-field assignments to allocate papers of a given author to a field of science. Then, we assign authors to the most common field listed in their publications. But each paper is assigned to the field that represents majority authors, and the audience for the paper. This is a major step. It replaces crude assignment of publications according to most common field for a journal with a more nuanced assignment that reflects interests of individual authors. To see the advantage, consider interdisciplinary journals. An article in **Science** or **Nature** written by a group of physicists is classified under our scheme as a physics paper, whose audience is likely to be other physicists. This is a clear advance over earlier practice, which assigns articles in these journals to the majority field, which in **Science** or **Nature** is biology or medicine. The approach divides an article among fields according to the majority field of its authors, not the majority field of the journal[16].

C. Determining the Sequence of Institutional Affiliations

The previous section describes the disambiguation of scientists' names, the resolution of scientists' identities, and the assignment of scientific papers to individuals. Remember, though, that institutional affiliations are not identified except for single institution papers. The remaining problem is to assign disambiguated individuals to a sequence over time of institutional affiliations. This section provides an overview of the affiliation algorithm that makes this feasible.

Changing institutions requires costly expenditures of time, money, and disruptions to lifestyle. We know from previous work (Breschi and Lissoni, 2009) that most researchers move less than three times in their careers. We use this insight to produce an algorithm that creates research histories with a minimum number of changes in institution affiliation. It is also reasonable to assume that if an institution name

[16] In rare cases a tie exists in which the number of authors is the same across two different fields. In these cases, the paper is assigned equally to the two fields. For example, a paper written by two chemists and two physicists is assigned to both chemistry and physics.

appears frequently in publications for an individual, then it is more likely that the individual is affiliated with that institution. Accordingly, we choose the sequence of institutions that minimizes the number of institutional changes for a given individual, taking account of the frequency with which an institution appears in each year, assigning greater weight to higher frequency institutions[17]. Using Dijkstra's Algorithm (Dijkstra (1956); see "Dijkstra's Algorithm" in **Wikipedia**) we create and assess an algorithm based on Dijkstra's approach. We check this algorithm using a hand-curated sample. We find that the method fits actual changes in institutional affiliations well for most researchers, defined as those who move rarely and are not afflicted by unusually large lumping or splitting errors. Even for researchers whose transition is incorrectly dated, aggregation of researchers at the level of nodes and years tends to cancel out errors of dating. For details see Appendix B of this paper.

III. Constructing the Networks

A. Nodes of the Networks

Building on Section II we collect scientists, papers, and citations received into institutions, locations, and fields by year[18]. Together the data cover the evolution of citation networks over 1981-2011.

Henceforth our discussion is concerned with nodes and edges of the networks. From prior steps individual scientists have been associated with an institution and field at different points in time.

Summing scientists by node and year yields the number of research scientists. Since we know prior years that a scientist publishes with a node, we can associate duration of publication at a node by a scientist. From this information we can construct numbers of scientists by node and duration of publication. We use this information to record research scientists by all durations of publication and by

[17] The data are too voluminous to identify institutional affiliations through manual checking.

[18] Since the data form networks, the number of entries for a single year is N^2 where N is the number of observations. We shall show that collecting our data into institutions, locations, and fields yields $N = 2{,}685$, so for a single year the number of entries is $N^2 = (2{,}685)^2 = 7{,}209{,}225$. Even this does not allow for linkages across years due to citations. The point is that even aggregating the data by institution, location, and field yields exceptionally large networks.

durations of six years or more, which we call long-term research scientists. By node and year, we also accumulate fractional and whole papers, and citations over the next five years to the papers,.

We consider permanence of a node to be a desirable trait. And yet, many nodes are small, publish occasionally, and only intermittently belong to networks. It is not clear that these are stable research groups. Across the sciences, we find that 1,793 U.S. university departments have graduate programs and receive federal funding. Since graduate teaching and pursuit of funding involve substantial fixed costs, such departments are likely to be permanent. The 1,793 departments capture all major research groups in U.S. universities.

Besides universities we identify 892 firms, government agencies and laboratories, and nonprofit biomedical centers for which, in each field, there are at least four hundred publications, and for which its locations produce at least one hundred publications. The number of 'real' nodes (institution-location-fields) is 2,685 (1,793+892), of which two-thirds are universities and one-third are other institutions. Not all these nodes exist at every point in time, so 2,685 is an upper bound on the number of nodes in any year. These restrictions are responsible for the limitation to 255 institutions noted in the introduction. But this set of institutions dominates U.S. scientific research.

B. Edges of the Networks

Consider edges that connect the nodes. In this study we are primarily interested in edges of citation networks in which nodes participate; only secondarily are we concerned with collaboration edges and collaboration networks. This is because it is difficult to distinguish separate effects of citation and collaboration centralities at the node level. See Section VI.C and Appendix D for more information.

Nevertheless, for purposes of clarification, we distinguish properties of the two networks. Citations are more abundant than collaborations. Citations result from literature search that is subject to constant marginal costs. Collaborations are costly; their marginal cost rises with distance between collaborators and the number of collaborations. Edges follow different rules in the two networks. Citations are controlled by citing scientists. Collaborations require the assent of all scientists in the team doing the

research underlying a paper. In citation networks edges are directed to the cited node from the citing node. Information flows in the reverse direction, from cited nodes to citing nodes. The citation network is a Directed Acyclic Graph (DAG) because citations from past to future are impossible[19]. In collaboration networks, edges are not directed, and the flow of knowledge is in both directions.

Papers contribute to edges between nodes of citation networks according to these rules. For each citation between a cited paper and a citing paper, a directed edge is added between each institution-location-field-year associated with the citing paper and each institution-location-field-year associated with the paper cited. The weight of each such edge is the inverse of the number of edges added to all pairs of nodes, so each citation contributes a total value of 1 to the citation network. An institution-location-field-year is the institution-location-field distinguished by year. Year is an important distinction because year cited precedes year citing[20].

As we have seen, to create manageable networks, we subdivided the networks into six-year intervals, where nodes are arranged by year citing. This results in six citation networks ranging over 1981-1986; 1986-1991; 1991-1996; 1996-2001; 2001-2006; and 2006-2011. Breaking up the data this way generates comparable time intervals and facilitates the study of network evolution. We sum citations across all citing and cited years falling within a time interval, between citing and cited institution-location-fields.

Now consider collaboration edges. If a paper has more than one author, a collaboration has taken place, and the paper contributes to the collaboration network for the year when it was published. Single author papers are omitted from the collaboration network, though they are not omitted from the evidence.

We consider three cases for papers having multiple authors:

1. The paper is associated with one institution-location-field: This implies an internal collaboration in the department. In this case an edge of weight one is added from the department to itself.

[19] Newman (2010), pages 118-121, shows that the eigenvalues κ of a Directed Acyclic Graph, such as a citation network, are identically zero. This sets the eigenvalue centrality to zero, making it useless for the analysis of citation networks. Since the eigenvalue centrality, x, is defined by $Ax = \kappa x$, where κ is the eigenvalue, only $x = 0$ satisfies the equation, given that $\kappa = 0$.

[20] We neglect rare examples where a citing paper references a paper published earlier that same year.

2. The number of authors is more than the number of institution-location-fields. This means that collaboration occurs between and within institution-location-fields. In this case an edge is added between each pair of departments associated with the paper (for between department collaboration) and from each department to itself (for within department collaboration). The weight on each such edge equals the inverse of the total number of edges added. This ensures that every paper contributes a value of one to the collaboration network.
3. The number of authors equals the number of institution-location-fields: This means that there is only one author per node. Hence, collaboration only occurs between institution-fields. In this case, an edge is added between each pair of institution-fields on the paper. Again, the weight on each such edge is set equal to the inverse of number of edges added.

We again divide collaboration networks into six-year intervals from 1981-1986, 1986-1991, through 2006-2011. As with citations, we sum collaborations over years belonging to the same interval, between any pair of institution-fields linked by collaboration.

For both citation and collaboration networks, the total weight placed on all edges is the sum over edges created by the above processes. It follows that a weighted edge is the sum of fractional cites or fractional collaborations between the institution-fields, since edges for a paper are fractional, as we previously explained in the construction of the weighted edges.

C. Citation Centralities

Creation of nodes and edges of the citation networks sets the stage for centralities needed to represent the influence of networks on nodes, more specifically on employment of research scientists in the nodes, as we shall show later in Section V. We are interested in the citation in-degree centrality and the citation PageRank centrality for weighted citation networks.

The weighted citation in-degree centrality for node i is defined as

(1) $$k_i^{in} = \sum_{j=1}^{n} w_{ij}$$

In equation (1) w_{ij} is the number of fractional citations from future node j to past node i, $w_{ij} \geq 0$. By fractional citations we mean citations distributed according to the fractions defined in the previous subsection, Section III.B. We also constructed the weighted collaboration degree centrality, and we compare its role to that of the citation in-degree in Section VI.C. These degree centralities are highly correlated, and their effects cannot be distinguished at the node level. The citation PageRank centrality is defined as

$$(2) \qquad x_i = \alpha \cdot \sum_{j=1}^{n} W_{ij} \frac{x_j}{d_j^{out}} + \beta$$

In equation (2), x_i is the weighted citation PageRank centrality for node i , α is a parameter usually set equal to 0.85, W_{ij} is the *row* of citation weights from future j to past i, x_j is the weighted citation PageRank centrality for node j, d_j^{out} is the weighted out-degree of j, and β is an additive constant, usually set equal to 1.0, and set equal to 1.0 in this paper. PageRank has the advantage that division by out-degree cancels the false transmission of centrality from more central nodes to less central nodes.

IV. Evolving Science Citation Networks in the U.S.

A. Average Degree Centrality and Network Density

Let us describe citation networks classified by fields of science. Using six-year intervals in Tables 1 and 2, we focus on changes between the first (1981-1986) and last (2006-2011) of the intervals. This yields twenty-five-year long differences in the network measures. An interesting point is that the density of the network, the actual number of edges divided by the maximum number of edges that is possible, is proportional to the average degree centrality (Newman (2010), pages 133-136)[21]. A rising average degree centrality, such as shown in Table 1, corresponds to a rising network density, and a rising density amounts to a more connected network.

[21] It is simple to extend Newman's proof to weighted networks. In the weighted case, the weighted density equals the weighted average degree centrality.

B. Changes in U.S. Science and Engineering Citation Networks

Tables 1 and 2 describe long-term changes in science citation networks between 1981-1986 and 2006-2011. Table 1 compares changes in network densities within fields, and in the bottom row, all fields. Edges equal 0 or 1 so densities are fractions. As shown in the first and second columns, densities rise uniformly over time. The bottom row, covering all fields, reveals a much smaller density, due to the low densities of edges between fields. But still the density for all fields rises three-fold.

For groups of sciences, Figure 1 graphs the increase in citation network densities by six-year intervals. This makes it easy to compare growth in densities over groups. Densities converge, and densities in mathematical sciences and engineering catch up to densities of other groups.

The third and fourth columns of Table 1 count changes in the number of components over time. Since a component is a subgraph that is not connected to the rest of the network, the number of components captures network fragmentation. Thus, a declining number of components over time implies that networks are becoming less fragmented. The number of components falls particularly sharply in engineering, which includes small, initially isolated groups in industry and government.

By necessity centralities are calculated on the largest component of a network, known as the giant component, which excludes observations in other components. This means that the centralities are computed on the largest component for each field and time interval. As time passes, the giant component encompasses more of the networks, but in the beginning, it is less inclusive.

Table 2 uses citation-weighted edges to study disciplinary orientation of citations. The four columns on the left cover 1981-1986, while the four on the right cover 2006-2011. Citation-weighted edges measure flows of ideas in the opposite direction. That is, citations from node B to node A imply a corresponding flow of ideas from A to B. The table compares flows of ideas inside fields to flows outside fields. Extending this to net flows, the net direction of citations between fields identifies the net direction of idea flows between fields. A minus sign in the fourth column on each side of the table implies a net inflow of ideas. For agriculture during 1981-1986, there are 14.2 thousand citations to agriculture from

other fields, but there are 16.2 thousand citations from agriculture to other fields. The net inflow of citations is negative, implying a net inflow of ideas to agriculture. Agriculture is therefore a net importer of ideas from other fields.

Citation-weighted edges increase over the twenty-five-year period. The rate of increase differs over fields, in part because of differences in the rate of increases in funding. Remarkably, citation-weighted edges to other fields, or interdisciplinary flows of knowledge, increase less rapidly than flows within fields. On average, fields of science are becoming slightly less interdisciplinary, judging by the bottom 'All Fields' row of the table.

Most fields cite more than they are cited. Agriculture, medicine, chemistry, earth science, and engineering fit this description. These fields are net importers of ideas and have a negative inflow of citations. A few fields cite less than they are cited. These fields are biology, mathematics and statistics, physics, atmospheric science, oceanography, and economics, which are net exporters of ideas. Only one field changes its status. This is computer science: it changes from being a net importer of ideas to a net exporter, consistent with the rise of the field in the practice of other fields. This concludes the discussion of Table 2.

C. Summary Remarks on U.S. Science and Engineering Collaboration Networks

The main focus of this paper is on citation networks, but to be systematic we record summary observations on collaboration networks[22]. Network collaboration density rises in all fields, and it rises faster in the collaboration network than in the citation network. Consequently, the average degree of a collaboration node rises faster than the average degree of a citation node.

As with citation networks, the number of isolated components in collaboration networks drops precipitously. Initially, there are more components in collaboration networks than in citation networks,

[22] Adams, Black, Clemmons, and Stephan (2005) find that the collaborations on scientific papers are growing larger and are more geographically dispersed over time, especially since the 1990s. Since the division of labor in science increases with team size, this suggests a rise in research productivity.

especially in engineering. But the number of isolated components in collaboration networks falls at a faster rate than in citation networks.

Figure 2 shows the increase in collaboration network densities over six-year intervals by groups of sciences. This allows us to compare growth in collaboration densities across groups. Apart from engineering, densities are converging. While engineering network density rises, it remains lower than in other groups of sciences. If interdisciplinary research is more important in engineering, these costs of interdisciplinary research could pose an entry barrier to rising network density.

D. Weighted Densities of the Science Networks

Since a rising weighted density implies that a node on average has more weighted edges attached to it and a higher degree centrality, this suggests greater efficiency and a more elaborate division of labor in U.S. science networks.

The changing costs of scientific research help to explain changes in network density. Costs of citation have fallen owing to growth of the Internet and the invention of efficient search engines. Costs of collaboration have fallen because of the Internet and improvements in Email and messaging. Federal policies also play a role by increasing the amount of funding, especially in favored fields of research.

For instance, federal defense R&D cyclically alters the density of parts of the networks that are intensive in defense research. Exogenous fluctuations in defense readiness introduce cycles in intramural defense R&D and in federal support of R&D in industry[23]. Most federal support of industry R&D is related to defense. In fact, half of it supports research in aircraft and missiles. Figures 3A and 3B present

[23] In our data the list of intramural defense institutions and defense-related federal laboratories is: US Department of Defense, Department of the Airforce, Department of the Army, Department of the Navy; NASA Ames Research Center, NASA Glenn Research Center, NASA Langley Research Center; Johns Hopkins University Applied Physics Laboratory; Lawrence Livermore National Laboratory, Los Alamos National Laboratory, MIT Lincoln Laboratory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, and Sandia National Laboratory. Defense contractor firms in aircraft and missiles include AlliedSignal Inc, Boeing Co, General Dynamics Corp, Goodrich (BF) Corp, Lockheed Martin Corp, Northrop Grumman Corp, TRW Inc, Textron Inc, and United Technologies Corp. Defense contractors in instruments include Honeywell Inc, Litton Industries, and Raytheon Co. Defense contractors in other manufacturing include Harris Corp, Hercules Inc, Motorola Inc, Rockwell International Corp, and Texas Instruments Inc. Altogether thirty-one institutions undertake substantial research associated with defense.

evidence from 1981-2011 on this point. Figure 3A plots the course of federally funded R&D in industry in millions of 2009 dollars. During the defense buildup of the 1980s, the flow of federally funded R&D rises from thirty billion to fifty billion dollars in 1987. Reflecting the 1991 breakup of the Soviet Union, the annual flow of federally funded R&D declines to twenty billion by 2001. With the onset of the War on Terror, federal support of R&D in industry reaches a peak of forty billion dollars in 2011. Manufacturing accounts for most of Federal R&D in industry during these years.

From 1987 on, federal intramural defense R&D falls from a peak of fifteen billion in 1987 to under ten billion by 2001. With the threat of terrorism, it thereafter rises to nineteen billion by 2011. And while federal support of R&D in industry never recovers its 1987 peak, federal intramural defense R&D undergoes a much milder cycle, growing to new heights by 2011. The recovery of intramural R&D is not enough, however, to prevent an overall decline in scientific research for defense.

To make changes in defense R&D policy transparent, Figure 3B normalizes the data by 1981 values. Despite the decline in R&D following the fall of the Soviet Union, trend growth prevails in defense intramural R&D, but defense R&D in industry a downward trend that is only partially abated after 2001 by the War on Terror.

Figures 4A and 4B take a closer look at federal support of R&D in industry. Years covered are 1981-2007[24]. Figure 4A graphs federal support in aircraft and missiles, instruments, and other manufacturing (principally electronics). These sectors account for the largest part of federal support in industry. Aircraft and missiles show a decline from a peak of thirty billion 2009 dollars in 1987 down to five billion in 2007. Federal R&D in other manufacturing also declines, though the decline is less severe. Only federal support of R&D in instruments increases, eventually becoming the leading sector in federal support of industry R&D[25]. Figure 4B normalizes the data by 1981 values, showing the compositional

[24] The NSF Survey of Business R&D was redesigned in 2008. This makes data from 2008-2011 hard to compare with data covering the years 1981-2007. Therefore, the graph ends in 2007. We thank the National Science Foundation for pointing out the reason for this change in the series.

[25] The instruments industry consists of search, detection, navigation, guidance, aeronautical and nautical system and instrument manufacturing, instrument manufacturing for measuring and testing electricity, and analytical laboratory instrument manufacturing.

shift underway in defense support of industry R&D, towards instruments and away from aircraft, missiles, and other manufacturing. We put the defense cycle to use in the regression analysis of Section VI.

Figures 3 and 4 reveal cycles in defense R&D spending. In contrast, Figure 5 shows sustained growth in federal support of university R&D. To highlight differences among groups of fields, we break Figure 5 into three parts. Figure 5A compares flows of R&D in millions of 2009$ in six groups of sciences. Life sciences dominate the figure, where federal support rises from four billion dollars in 1981 to twenty billion in 2011 in 2009 dollars. Figure 5B excludes life sciences. This makes it easier to observe growth of federal support of other fields. Physical science and engineering stand out in Figure 5B. To render growth in all fields visible, Figure 5C normalizes every series by 1981 values. When this is done, federal support of university R&D in mathematical sciences clearly grows fastest, reflecting growth of computer science R&D. The second fastest growing is engineering. We conclude that federal support grows fastest in computer science and engineering, consistent with a perception of their rising value to the U.S. economy. Only trend differs: all fields show an increase in support. Because of this, trends in federal funding are best suited as instruments for node level research funding in the regressions of Section VI.

V. Behavior of Nodes of Science and Engineering Networks

Scientists are employed by research institutions in their previously chosen field of specialization, so occupational specialization is predetermined. Contingent on occupation, the key question is the allocation of the scientist's time and effort over tasks. For simplicity we assume two tasks, administration (or administration/teaching) and research. The scientist chooses the mix of administration and research that maximizes income.

The literature of occupational choice by Roy (1951), Borjas (1987) and others emphasizes the role of comparative advantage. Our focus is instead on maximization of returns to a given occupation by choice over tasks, following Autor, Levy and Murnane (2003) and other papers in the task-based literature. Adhering to Becker (1985), scientists maximize income subject to time and effort constraints.

Administration may be the primary task for most scientists due to a greater value placed on administration and the greater importance of administrative skills in the population.

A. Choice by Individual Scientists

Scientist i owns stocks of effort and administrative and research skills, which are denoted by E_i, A_i, and R_i. As noted, the problem is one of allocating E_i, A_i, and R_i among administration and research tasks to maximize income. The allocation of effort is represented by E_{Ai} and $E_{Ri} = E_i - E_{Ai}$. Given one unit of labor supplied inelastically, the allocation of time is expressed by a fraction λ_i allocated to administration and $1 - \lambda_i$ is allocated to research. Earnings Y_i consists of income from tasks based on a constant elasticity function, connected by the effort and time constraints:

$$(3) \qquad Y_i = P_A \cdot A_i \cdot E_{Ri}^{\mu} \cdot \lambda_i^{\sigma} + P_R \cdot R_i \cdot (E_i - E_{Ai})^{\mu} \cdot (1 - \lambda_i)^{\sigma}$$

Prices P_A and P_R convert task output into value of task output. Shocks to task prices shift the allocation from one task to the other. To illustrate, a negative (or positive) shock to the administrative price would redirect the scientist into (or out of) research and similarly for the research price.

Assume that the market assigns a higher price to administration so $P_A > P_R$, and assume that administrative skills are greater: $A_i > R_i$. These assumptions favor administration over research, perhaps implying specialization, if $\frac{\partial Y_i}{\partial E_{Ai}} > 0$ and $\frac{\partial Y_i}{\partial \lambda_i} > 0$ at the point of specialization. In this case scientists will not do research and even for those who do, most are part-time researchers. Employment of (fractional) research scientists responds to a shock in research productivity favoring a reallocation towards research. Research scientists are a proportion of staff that responds to research productivity, and research scientists are the focus of this paper.

Consider an interior solution where the scientist does research as well as administration. Tasks are subject to diminishing returns, since $0 < \mu < 1, 0 < \sigma < 1$, and $0 < \mu + \sigma < 1$. Maximizing (3) over E_{Ai} and λ_i yields the first order conditions:

(4) $$E_{Ai} = \lambda_i = \frac{(a_i)^{\frac{1}{1-\mu-\sigma}} E_i}{1+(a_i)^{\frac{1}{1-\mu-\sigma}}}$$

(5) $$E_{Ri} = 1 - \lambda_i = \frac{E_i}{1+(a_i)^{\frac{1}{1-\mu-\sigma}}}$$

Appendix C proves these results. In (4) and (5) $a_i \equiv P_A A_i / P_R R_i$ is the ratio of values of administration and research. It measures comparative advantage in administration. Equations (4) and (5) indicate that an increase in a_i increases administrative effort E_{Ai} and reduces research effort E_{Ri}. An increase in the stock of energy E_i increases E_{Ai} and E_{Ri} in the same proportion, since more energetic scientists devote more energy equally to administration and research[26].

The equal effect of administrative comparative advantage a_i is the result of the equal elasticities assumed of administrative and research tasks in the earnings function (3)[27]. While (4) and (5) yield the conclusion that scientists follow task-based comparative advantage by allocating more effort and time to administration as a_i increases, they will not specialize as readily in research as they would in a linear model of skills (Autor, Levy, and Murnane, 2003), because returns to effort and time diminish.

Since scientists rarely specialize in research the labor force in administration and research is the sum of often *fractional* workers in each activity. Furthermore, if comparative advantage of administration (a_i) increases, the research labor force is absorbed into the administrative labor force. Conversely, the research labor force absorbs part of the administrative labor force when the research price P_R rises relative to the administrative price P_A. The boundary between administration and research is flexible and determined by comparative advantage in tasks[28]. Another implication is that effective hours are increased

[26] The prominent early mathematical economist R.G.D Allen appears to have been equally at home in administration as in research at the London School of Economics.

[27] Though a closed form solution is not available when elasticities of output differ over administration and research, one can calculate the displacement system around the point of equal elasticities when slight changes in the elasticities are allowed for obtaining results for the more general case.

[28] This flexible supply of research scientists argues that the discontinuous labor adjustment costs emphasized by Hamermesh (1989) and Caballero, Engel and Haltiwanger (1997) are less important, since fixed costs of training are less relevant given that research scientists can reallocate towards research if it is in their interest to do so.

by the stock of effort E_i so the scientist has an incentive to increase the stock by proper diet, exercise, rest, and limited television watching. The model also implies that, since participation in networks increases available R_i by teamwork and by expanded resources, effort and time aimed at research increase. We return to these points in the discussion of the institution, location and field node below.

B. Linear-Quadratic Model of a Node

To understand the behavior of a node—an institution, location, and field—we employ a version of the linear-quadratic model of social interactions. This is a core model in the social and economic network literature (Ballester, Calvó-Armengol, and Zenou, 2006; Jackson (2008), Calvó-Armengol, Patacchini, and Zenou, 2009; Bramoullé, Kranton, and D'Amours, 2014; Acemoglu, García-Jimeno, and Robinson, 2015; Blume, Brock, Durlauf, and Joyaraman, 2015).

In our version of the linear-quadratic model, a node maximizes the value of administration and research over administrative and research scientists in the labor force. Recall that individual scientists choose the intensive margin between these activities to maximize earnings. The node adjusts along the extensive margin by changing fractional numbers of scientists in each task.

A node is assumed to maximize the sum of the value of administration and research minus costs of production. To a first approximation, we treat the returns and costs as separable in administrative and research scientists[29]. This allows us to study research independently from administration.

We assume that research is subject to more extensive network effects than administration. We make this assumption because research clearly benefits more from meetings with other nodes, from knowledge flows due to citations, and from external collaborations, in a way that administration does not. These interactions are represented in our implementation.

[29] This is true for the prevailing case of undergraduate education in universities. For graduate education we can allow for interactions between instruction and research and maximize $NV_{Ai} + NV_{Ri}$. But even in this case the cross effects of instruction on research can go either way. We set the interaction approximately to zero.

Building on the linear-quadratic formulation in Acemoglu, García-Jimeno, and Robinson (2015), and assuming separability of returns and costs in administration and research, as well as asymmetric network effects on research, we write the value of administrative output $p_{Ai}q_{Ai}$ as a linear-quadratic function of administrative scientists ℓ_{Ai}, and likewise for the value of research output $p_{Ri}q_{Ri}$ and research scientists ℓ_{Ri} . Prices and quantities of administration and research are p_{Ai} and p_{Ri} and q_{Ai} and q_{Ri} respectively. The node's problem is:

$$(6) \qquad \underset{\ell_{Ai},\ell_{Ri}}{max} \begin{Bmatrix} NV_{Ai} \equiv p_{Ai}q_{Ai} - \frac{\eta_A}{2}\ell_{Ai}^2 = v'_{Ai}\varphi_A + \beta_{Ai}\ell_{Ai} - \frac{\eta_A}{2}\ell_{Ai}^2 + e_{Ai} \\ NV_{Ri} \equiv p_{Ri}q_{Ri} - \frac{\eta_R}{2}\ell_{Ri}^2 = v'_{Ri}\varphi_R + \beta_{Ri}\ell_{Ri} + \gamma_C\left(\sum_j C_{ji}\ell_{Rj}\right)\ell_{Ri} \\ +\gamma_T\left(\sum_j T_{ji}\ell_{Rj}\right)\ell_{Ri} - \frac{\eta_R}{2}\ell_{Ri}^2 + e_{Ri} \end{Bmatrix}$$

The terms β_{Ai} and β_{Ri} are random coefficients given by:

$$(7) \qquad \beta_{Ai} = x'_{Ai}\theta_A + u_{Ai}$$

$$(8) \qquad \beta_{Ri} = x'_{Ri}\theta_R + u_{Ri}$$

From equation (6), $p_{Ai}q_{Ai} = v'_{Ai}\varphi_A$ and $p_{Ri}q_{Ri} = v'_{Ri}\varphi_R$ when $\ell_{Ri} = \ell_{Ai} = 0$. The node produces presumably negligible, serendipitous output without labor, and all other terms must interact with administrative or research scientists. In particular, network effects in the bottom equation of (6) for research scientists, must interact with these scientists[30]. Therefore, following Manski (1993), network effects are exogenous or contextual effects that interact with network centralities.

In the bottom equation of (6), we include network terms$\left(\sum_j C_{ji}\ell_{Rj}\right)$ and $\left(\sum_j T_{ji}\ell_{Rj}\right)$, which are rows of the citation and collaboration networks pointing at department i. First, the department is cited by other departments, as shown by $\sum_j C_{ji}\ell_{Rj}$, giving rise to more resources for i. Second, it collaborates with other departments, as shown by $\sum_j T_{ji}\ell_{Rj}$. From now on, we simplify by writing the network terms as

[30] Migration to and from node i is dealt with during the construction of research scientist counts ℓ_{Ri} in Section III.A.

centralities $C_{Ci} \equiv \sum_j C_{ji}\ell_{Rj}$ and $C_{Ti} \equiv \sum_j T_{ji}\ell_{Rj}$. Here C_{Ci} is the citation centrality and C_{Ti} is the collaboration centrality.

The top equation of (6), for administration, lacks network terms. It conforms to our view that knowledge flows and resource sharing are important for research but not administration. However, networks of contacts could be keys to administration. Alternatively, research networks could enter indirectly if they lead to reallocation from administration to research. In this case, network effects could appear with a negative sign in the administrative net value equation.

First order conditions for ℓ_{Ri} are:

(9) $$x'_{Ri}\theta_R - \eta_R\ell_{Ri} + \gamma_C C_{Ci} + \gamma_T C_{Ti} + u_{Ri} = 0$$

Second order conditions are satisfied since η_R is positive. Solving (9) for the best-response demand equations for research scientists (Acemoglu, García-Jimeno, and Robinson (2015)) yields:

(10) $$\ell_{Ri} = \frac{1}{\eta_R}[x'_{Ri}\theta_R + \gamma_R C_{Ci} + \gamma_T C_{Ti} + u_{Ri}]$$

Equation (10) indicates that the citation and collaboration centralities increase (decrease) employment of research scientists if $\gamma_R > 0$ and $\gamma_T > 0$ ($\gamma_R < 0$ and $\gamma_T < 0$), in which case the centralities are strategic complements (substitutes). Inserting (10) into the research value equation at the bottom of (6) yields

(11) $$p_{Ri}q_{Ri} = v'_{Ri}\varphi_R + \frac{\eta_R}{2}\ell_{Ri}^2 + e_{Ri}$$

Since p_{Ri} is unknown we normalize (11) by output price and set $p_{Ri} = 1$. The empirical work concentrates on the best response demand equation (10) for research scientists, and the optimal supply of research output (11).

In summary, the implications of this section are:

1. To a first approximation, network effects are concentrated on research scientists.

2. In theory, there are two network effects, one from citation, the other from collaboration, though in practice a separate effect of collaboration is hard to distinguish owing to collinearity of the centralities.
3. If strategic complementarity prevails, network effects increase research scientists on the extensive margin of the institution, location, and field.
4. If there is a behavioral response from individual scientists, network effects lead to a reallocation from administration to research on the intensive margin. In that case network effects could enter negatively into the equation for administrative scientists if administrative labor were observable.

VI. Econometric Findings

The best response demand for research scientists and supply of research output are the equations to be estimated. At issue is whether the centralities increase or decrease employment of researchers, and are therefore strategic complements or substitutes. Assuming that citation and collaboration centralities are collinear, as we find (see Appendix D), we concentrate on citation centralities. We estimate the best response of research scientists to citation in-degree centralities or citation PageRank centralities because of the useful properties of these centrality measures. See Section IV. However, endogeneity problems result in biased centrality coefficients. Omitted attributes of nodes are correlated with the centralities, resulting in a correlation between the centrality and the error term in the best response demand equation. We deal with the problem by random assignment of centralities from another node in the same field and institutional type, but in a previous time interval and hence a different network. This instrument breaks the link between omitted variables and the centralities and eliminates the bias in the centrality coefficient.

Part A. of this section describes variables used in the regression analysis. Part B clarifies the randomly assigned instruments for the centralities. Two-stage least squares (2SLS) estimates of the best response equations for research scientists are presented in Part C. The section concludes with Part D., which gives details of three-stage least squares (3SLS), two equation estimates of the jointly estimated, best response demand for research scientists and research output, where 3SLS is generalized to include errors clustered by institution, location, and field. Tables 11A and 11B contain these results.

A. Description of the Covariates

Table 3 reports descriptive statistics for the covariates: the number of observations, mean, standard deviation, minimum, and maximum. The panel of nodes contains 77,048 observations during 1981-2011. Recalling, though, that centralities are necessarily calculated on the largest, or giant component of the networks, the centralities are estimated on 59,862 observations. This shows that the giant component is about seventy-eight percent of the sample of 77,048 observations. The sample size reflects a sixty-paper cutoff during each six-year interval. We have applied a thirty-paper cutoff that yields a larger sample, but with similar regression results. We also include citations received from other institutions over the next three and five years from other institutions. Self-citations are excluded.

The top ten lines of Table 3 are concerned with the best response equation. Line one describes all research scientists over all different tenures with a node. This includes transient researchers and researchers who publish intermittently, as well as long-term researchers. The line two describes long-term research scientists. They are defined as steadily publishing researchers who appear for at least six consecutive years in the data for a node. The mean number of all research scientists is 261, while the mean number of long-term research scientists is 180. Long-term research scientists are the majority except for top departments, indicating a substantial number of graduate students and visitors in these nodes. The maximum number of researchers is in the tens of thousands. This takes place in top medical schools, and it is due to collaborations with research scientists employed elsewhere, such as physicians in private practice or faculty in smaller institutions. Top medical schools have many outside collaborators. Our method makes observable the 'invisible college' around these schools.

Lines three and four report citation in-degree and PageRank centralities. The in-degree centrality has a mean of 1,687.54. To get this result, notice that there are about one hundred papers per node, which are cited roughly sixteen times. The table shows how different the PageRank centrality is from the in-degree centrality. PageRank centrality averages 0.01. This is because the centrality is normalized. It sums to 1.0

over all nodes in a field. The average number of nodes per field is about one hundred, thus accounting for this result.

Line five contains an indicator variable for employment in other fields in the same location and institution. This probably stands for larger nodes, since small nodes outside universities employ research scientists in one field. These nodes are very few: single field nodes account for just 1.5 percent of the observations. Line six contains an indicator variable for employment in the same field in other locations for the same institution. Since universities represent one location in the National Science Foundation data, multiple location institutions are firms or government and nonprofit laboratories. This variable could indicate substitution between locations. It occurs in about 18 percent of the observations.

Lines seven and eight measure exposure to the defense R&D cycle. Defense Intramural Dummy is an indicator variable for defense agencies and federal laboratories engaged in defense research[31]. Defense Intramural R&D (1981=1.0) is national intramural defense R&D in millions of 2009 dollars, normalized by the 1981 value. The intramural defense cycle variable is the product of the two variables noted above:

Defense Intramural Dummy × Defense Intramural R&D (1981=1.0)

Similarly, Defense Contractor Dummy is an indicator variable for firms that undertake substantial defense research[32]. Defense Contractor R&D (1981=1.0) is national defense contractor R&D, expressed in millions of 2009 dollars, normalized by the 1981 value. The contractor defense cycle variable is the product of the two variables noted above:

Defense Contractor Dummy × Defense Contractor R&D (1981=1.0)

[31] Fourteen institutions are defense intramural institutions: US DEPT DEFENSE, US AIR FORCE, US ARMY, US NAVY, AMES RES CTR, GLENN RES CTR, LANGLEY RES CTR, JHU APPLIED PHYSICS LAB, LAWRENCE LIVERMORE NATL LAB, LOS ALAMOS NATL LAB, MIT LINCOLN LAB, OAK RIDGE NATL LAB, PACIFIC NORTHWEST NATL LAB, and SANDIA NATL LAB.

[32] Seventeen institutions are defense contractors: BOEING CO, GENERAL DYNAMICS CORP, LOCKHEED MARTIN CORP, NORTHROP GRUMMAN CORP, RAYTHEON CO, TRW INC, TEXAS INSTRUMENTS INC, MOTOROLA INC, GOODRICH (BF) CO, UNITED TECHNOLOGIES CORP, ROCKWELL INTL CORP, LITTON INDUSTRIES INC, HARRIS CORP, HONEYWELL INC, ALLIEDSIGNAL INC, TEXTRON INC, and HERCULES INC.

As shown in Figure 3B, contractor R&D declines relative to intramural R&D over time. Note that for non-defense institutions defense cycle variables equal zero.

Lines nine and ten of Table 3 complete coverage of the best response demand equation. Line nine reports average distance to top research cities. Line ten reports the number of research scientists elsewhere in the same Metropolitan Statistical Area (MSA). Average distance to leading research cities is 1,163 miles[33]. Learning about innovative research could be more costly at greater distances. This could weaken incentives to employ research scientists and produce high impact research.

Average research scientists elsewhere in an MSA equal 307. To obtain this variable we divide local research scientists by the square root of distance from the institution in question. Normalization by square root of distance indicates rising costs due to distance. The meaning of research scientists in an MSA is unclear. The benign view suggests informal meetings where local scientists share ideas, implying complementary between local nodes. The cynical view is that local research scientists in an MSA compete for local research dollars, implying substitution between local nodes.

Table 3 concludes with variables the research output equation. The dependent variable is the number of citations received over the next five years. The main independent variables, following equation (11), are squared values of research scientists and fixed effects for institution, field, and year. On average, departments receive 491 citations over the next five years. The squared value of all research scientists is 575,555 and the squared value of long-term research scientists is 332,459.

B. Discussion of Centrality Instruments

Table 4 explains the instrumental variables for the centralities. For each centrality that we consider (in-degree centrality, PageRank centrality), the instrument is lagged by one or two six-year intervals. It is a centrality that is randomly assigned to a node from another node in the same field and institutional type,

[33] Top research cities are Boston, New York, Philadelphia, Washington-Baltimore, Chicago, San Francisco, and Los Angeles.

or else it is a centrality from the field most cited by the focal field, but again from the same institutional type. Examples of most cited fields are physics for electrical engineering and biology for medicine.

Consequently, there are four kinds of centrality instruments: same field versus most cited other field, and one-period lag versus two-period lag. The type of institution is the same type. By this we mean that random assignment is to one university from another university, and likewise for the other institutional types described previously. We impose the restriction of same type because size of these diverse institutions varies widely, producing weak instruments across types.

For each instrument we set a seed using the last six digits on a treasury bill. Having set the seed we define a population of centralities from a lagged network that excludes the citing institution but includes other institutions in the target population. To repeat, this is either the same field and same type, or the most cited field and same type[34]. We draw a uniform random sample of one for each citing institution and field to define the instrument for the observed centrality. Table 4 explains. Resulting instruments for the centralities exhibit a much lower correlation with research scientists on the left of the best response demand equation: see (10) in Section V.B. For example, the correlation between the in-degree centrality and all research scientists in a node is 0.88, while the correlation between the randomly assigned instrument for the in-degree centrality and all research scientists is 0.26. This suggests that we have removed much of the correlation between the error term and the centralities.

C. Collinearity Between Citation and Collaboration Centralities

Section IV covered citation networks but it only briefly covered collaboration networks, despite the inherent interest of both. From equation (10) we would seem to want both citation and collaboration centralities in the best response demand equations. Unfortunately, the two centralities are highly

[34] For example, the Stata statements are: "set seed 148983; sample 1, count by (node_name field);". The purpose of the set seed command is to create a particular reproducible sequence for the sample. The purpose of the sample 1 command is to select another institution by its ID (node_name) and, in this example, the same field of science. Previous code removes same institution from the sample, so self-citation is ruled out.

correlated, and their effects cannot be distinguished[35]. This makes sense, because highly cited departments also attract many collaborators. For example, when we include the citation in-degree centrality in equation 5.1 of Table 5, its coefficient, with all research scientists as the dependent variable, is 0.137, and the R-squared is 0.9268. If we were to add the collaboration degree centrality to this specification, the coefficient on citation centrality falls to 0.114 and the coefficient of the collaboration centrality is 0.533. The R-squared is now 0.9297 but the explanatory power of the regression is almost the same (0.9297 versus 0.9268). Adding the collaboration centrality does not change the explanatory power of the regression. Now repeat the experiment using equation 5.4 of Table 5, where long-term research scientists are the dependent variable. The coefficient of the citation in-degree centrality is 0.085 and R-squared is 0.8438. If we were to add the collaboration degree centrality, the coefficient on the citation centrality is 0.084, R-squared is 0.8440. The coefficient in the collaboration centrality is 0.012 and it is statistically insignificant. Adding the collaboration degree centralities fails to increase the explanatory power of the regressions. Hereafter, we confine the estimation procedures to the citation centralities, though for the sake of transparency we report expanded regressions including both centralities in the online Appendix D.

D. Estimates of Best Response Demand Equations

Table 5 presents OLS and 2SLS estimates of the best response demand equations for research scientists using the citation in-degree centrality. We report separate estimates for all research scientists on the left and long-term research scientists on the right of the table. All equations include fixed effects for institution, field, and six-year period. Table 5 assumes robust errors or else errors clustered by node. Randomly assigned instruments for the centralities are taken from the network one period before. For example, the 2006-2011 network draws instruments from the 2001-2006 network.

[35] The correlation between the citation in-degree centrality and the collaboration degree centrality is 0.929. The correlation between the citation PageRank centrality and the collaboration PageRank centrality is 0.843.

Equations 5.1 and 5.4 are benchmark OLS results. The coefficient of the citation in-degree centrality is positive and significant. It is somewhat larger for all research scientists (0.137) than for long-term research scientists (0.085). This is because high centrality nodes employ a larger share of short-term research scientists. Equations 5.2-5.3 and 5.4-5.5 are 2SLS results where robust or clustered standard errors are assumed. The coefficient on in-degree centrality stays about the same, but as expected the 2SLS standard errors increase substantially.

The indicator variable for other fields has a positive sign. This implies that multiple field nodes employ more scientists. The indicator for other locations has a negative sign. Multiple location nodes employ fewer scientists, perhaps due to substitution between locations. The defense cycle variables are insignificant in the all-research scientist equations 5.1-5.3, but they are positive and usually significant in the long-term research scientist equations 5.4-5.6. During the contraction phase of the defense cycle, long-term research scientists decline due to separations from the employer. During the expansion phase of the cycle, long-term employment recovers. The all-scientists variable conceals this effect, because short-term research scientists move in the opposite direction from long-term research scientists. During a contraction, long-term researchers leave defense institutions and are replaced by short-term researchers, and conversely for an expansion.

Average distance to leading research cities has a negative effect on research scientists. Perhaps innovative research is more costly to perform at a greater distance, though the negative effect of distance is not always significant in later tables. The negative coefficient of research scientists elsewhere in the same MSA suggests that funding is crowded out. But here too the negative sign disappears in some tables.

To conclude Table 5, we find that the excluded instruments are significant in the first stage equation. Tests of the overidentifying restrictions accept the null hypothesis that excluded instruments are valid.

Table 6 explores differences in the coefficient of in-degree centrality by type of institution. It interacts dummies for type of non-university institution with the centrality and enters the interaction as a separate endogenous variable. The four types of institution are defense institutions, government institutions, firms, and biomedical research centers. The specification follows equation 5.2 of Table 5 except for the added

interaction term. Table 6 is an excerpt that concentrates on coefficients and standard errors for the main centrality term and the interaction of the institution-type dummy with the centrality. Since we use an instrument for the centrality and the interaction of the institution-type dummy with the instrument for the centrality, we are using the instrument in a nonlinear setting twice. For this reason, we present tests for weak instruments. The tests reject the hypothesis that instruments are weak. The minimum eigenvalue statistics exceed even stringent critical values for H0, usually by a wide margin. Panel A compares each type of institution with the rest of the sample, while Panel B compares each type of institution with universities. For defense and government institutions the centrality interactions are insignificant. Strategic complementarity is the same for these institutions as for the rest of the sample or universities. Centrality interactions are negative for firms and at times biomedical research centers. While strategic complementarity is weaker in these cases, it is still positive and significant.

Using Panel IV fixed effects regression, Table 7 emphasizes the within dimension of the panel. Since centralities are used to compare the importance of different nodes, one might think that the centralities would be insignificant. With this thought in mind, the table examines whether the citation in-degree centrality retains any significance in the best response equations. Surprisingly, while the coefficient of the centrality falls, it remains significant in equation 6.1, indicating some role for the centrality in the within dimension, in comparing a node with itself, by means of centralities that vary over six-year periods.

Table 8 replaces the in-degree centrality with the PageRank centrality. Otherwise, Table 8 repeats Table 5. All research scientists are the dependent variable in 8.1-8.3, while long-term research scientists are the dependent variable in 8.4-8.6. Equations 8.1 and 8.4 contain benchmark OLS results. 2SLS results are found in 8.2-8.3 and 8.5-8.6. Coefficients of the other covariates differ slightly because PageRank replaces the in-degree centrality, but the coefficient of interest is the coefficient of PageRank. Recall that PageRank sums to 1.0 across nodes. Since the average PageRank centrality is 0.01, its coefficient is many times larger than the coefficient of the in-degree centrality. The crucial point, though, is that like the in-degree centrality, the coefficient of PageRank is positive and significant. Through the centrality, research

in one node stimulates employment of research scientists in other nodes. This again defines a strategic complement.

Table 9 increases the lag on the centrality instruments utilized in Table 5 from one six-year period to two periods. For instance, the 2006-2011 network draws on the 1996-2001 network for its centrality instruments. The coefficient on the citation in-degree centrality stays about the same as in Table 5. The coefficients of the other covariates are also about the same.

Table 10 employs two period lags as in Table 9, but it replaces the in-degree centrality with the PageRank centrality. A good comparison is with Table 8, which also employs PageRank centralities, even if with a one-period lag. The coefficient on the PageRank centrality roughly doubles its value in Table 8. It is positive and significant at the five percent level or above. Coefficients of the other variables are as before. This ends the discussion of single equation estimates of best response demands for research scientists.

E. System Estimates of the Demand for Scientists and the Supply of Research

This section reports system estimates of the demand for research scientists and the supply of research output using equations (10) and (11), which are reproduced below for convenience:

$$(10) \qquad \ell_{Ri} = \frac{1}{\eta_R}[x'_{Ri}\theta_R + \gamma_R C_{Ci} + \gamma_T C_{Ti} + u_{Ri}]$$

$$(11) \qquad q_{Ri} = v'_{Ri}\varphi_R + \frac{\eta_R}{2}\ell^2_{Ri} + e_{Ri}$$

More information is needed for (11): a measure of research output q_{Ri} (with $p_{Ri} = 1$) and an instrument for research scientists squared (ℓ^2_{Ri}), since scientists are correlated with the error term e_{Ri}.

We choose fractional citations received by a node over the next five years to measure research output. By fractional citations we mean citations adjusted by the authors resident in a node out of all authors listed on papers associated with the node. Taking account of future impact eliminates years 2007-2011 given that future citations are not observed after 2006.

If it could be found, an exogenous counterpart to past federal R&D funding should be highly correlated with research scientists. It could be a compelling instrument, because federal money is the most important source of support for research scientists in universities. Figures 5A-5C have shown us that growth of federal R&D differs among fields as a matter of policy. We know that federal R&D is concentrated among top university departments. And finally, we know that research scientists, including many working outside universities, receive federal support through faculty grants in universities. These pieces of information factor into the first funding instrument.

We compute this instrument as follows. For universities, which form two-thirds of our observations, we construct average federal R&D funding in millions of 2009 dollars during the first three years of the data series for a university and field. The data come from the National Science Foundation Survey of Higher Education R&D (known as HERD), which begins in 1973. For many fields, the first three years are 1973-1975. In detailed engineering and environmental sciences, the first three years are 1980-1982 (materials science is the exception: the data begin in 1990). Since the data in the rest of this study date from 1986 and later, because lagged instruments for the centralities eliminate the period 1981-1986, and sometimes the period 1986-1991 as well, the first three years of funding can precede the period of analysis by as much as thirteen years.

We carry the first three years of funding forward through subsequent years using the national growth rate of funding in a field. Average funding over the first three years smooths the initial annual data and works as a seed for the time series of the funding instrument. The average picks up advantages of top departments over others. But thereafter, the variable follows science policy by imposing the national growth rate of federal funding in a field.

In equation form, initial federal R&D funding in field i and university j in millions of 2009 $ is the average over years one to three:

$$\overline{RD}_{ij} = \sum_{t=1}^{3} RD_{ijt} /3 \qquad (12)$$

Thereafter the *simulated* flow of R&D grows according to the national growth rate in the field. Thus,

(13) $$RD_{ijt}^{Sim} = \overline{RD}_{ij}\left(1 + g_{32}^{i}\right) \cdot \left(1 + g_{43}^{i}\right) \cdot \ldots \cdot \left(1 + g_{t,t-1}^{i}\right)$$

In equation (13), $g_{t,t-1}^{i} = \left(RD_{t}^{i} - RD_{t-1}^{i}\right)/RD_{t-1}^{i}$. The growth rate is first applied to the second and third years of the funding series for a node, g_{32}^{i}. This is the national growth rate in the flow of R&D funding in field i between t and $t-1$.

The simulated flows in (13) are an interaction between the initial three-year average for the university and field and subsequent growth due to federal policies towards a field. The final variable, which is the instrument we use, is the simulated stock of federal R&D over the previous eight years:

(14) $$RDK_{ijt}^{Sim} = \sum_{l=1}^{8}(1-d)^{l-1} RD_{ij,t-l}^{Sim}$$

In (14) the stock is depreciated by a depreciation rate of $d = 0.15$. Equation (14) is the stock of R&D for a university and field. It is the university-field-time funding instrument.

For non-universities we lack initial R&D data that is specific to an institution and field, which provides the seed for the series (12) leading to the stock (14)[36]. Since research scientists in non-universities are mostly trained in universities, we impute the stock using depreciated flows of the national average R&D in a field *in universities* over the previous eight years:

(15) $$\overline{RDK}_{it} = \sum_{l=1}^{8}(1-d)^{l-1}\overline{RD}_{i,t-l}$$

Including imputes (14) and (15) deliver the first funding instrument.

The second funding instrument applies (15) to all observations for U.S. universities and non-universities alike. Note that (15) is a field-time instrument that is subject to far greater errors of measurement than (14). However, it is entirely the result of national funding in a field. It is a pure science policy instrument. Finally, we include the total MSA population in 1990 as an additional instrument, on the grounds that research scientists are more readily hired in larger populations.

[36] In the U.S., industry and federal R&D is not available by field of science and institution. This is probably because R&D in these sectors for the most part is not basic research and is not classifiable by science field.

Table 11A reports 3SLS results with clustered errors, using the citation in-degree centrality in the best response demand equation. For conciseness we limit reporting to the variables of interest, the centralities and research scientists squared. In Panel A the funding instrument is the university-field-time measure: (14) for universities and (15) for non-universities. As we have seen, it is an interaction between the mean of initial R&D over the first three years for the node, and the subsequent national rate of growth for the field. Centrality coefficients rise slightly compared to Table 5 and they remain positive and significant. Therefore, strategic complementarity prevails as well in these two equation system estimates. Besides this, the coefficient of the square of research scientists in the research output equation is positive and significant, indicating a rising supply of output[37].

Panel B uses the second funding instrument. It is the national average stock of R&D (15) for nodes in each field. While the centrality coefficients drop slightly, they remain positive and significant. Again, strategic complementarity applies.

Table 11B replaces the in-degree centrality with the PageRank centrality. Panel C employs the first funding instrument. This is (14), the interaction between the initial fixed effect for a node and national growth in the field for universities, and for non-universities, it is (15), the average stock in a field. The centrality coefficient estimates are higher than in Table 8. They are positive and significant. Panel D uses the average stock of R&D in a field as the funding instrument. In this case the centrality coefficient, while positive, is not significant at the ten percent level. This concludes the empirical work, which strongly supports the hypothesis of strategic complementarity between nodes of the networks, and which implies a supply curve of research output in the final tables, Tables 11A and 11B.

[37] The derivative of citations over the next five years with respect to research scientists is twice the coefficient times the number of research scientists. Using 0.000282 for all research scientists in Panel A, Table 10A, System 1, and recalling from Table 3 that the mean of all research scientists is 260.60, the derivative is $2 \times 0.000282 \times 260.60 = 0.1470$. An additional research scientist delivers 0.1470 citations over the next five years. The derivative for System 2, for a mean of 179.61 for long-term researchers, is $2 \times 0.000485 \times 179.61 = 0.1742$.

VII. Summary and Conclusions

This paper has presented findings on networks of scientists and engineers using data derived through a series of steps from the Web of Science™ database for the United States during 1981-2011. Based on citation networks that we have constructed covering major U.S. research institutions and fields of science, we find that the networks have grown more dense and less fragmented with time, and thus more effective. Citation weights on links between research institutions have increased, suggesting rising information flows between scientists in different organizations. A plausible reason—and a testable hypothesis—as to why this is happening is that science funding has expanded, especially in universities, driving up the volume of research.

Growing density and declining fragmentation are the result of three causes pulling in different directions. Federal support of academic science increases the density of the networks at different rates depending on growth of federally funded R&D by field of science. Partway into the period, the growth rate of federal R&D in computer science and engineering begins to grow more rapidly than federal R&D in other fields, driving the density of network connections upward in computer science and engineering. A second factor that affects density of network connections is the behavior of defense support of scientific research. This affects two sectors, intramural defense scientific research inside federal government, and defense-supported scientific research in defense contractor firms. The evidence presented here points primarily to a defense stand-down that lessens density of networks, in defense-related sectors, disproportionately affecting defense contractors. This factor has the opposite effect on density of network connections from the expansion of federal support of academic research. Finally, holding other factors constant, we find that density of the networks rises purely as a function of time. This is likely the result of hard-to-measure improvements in internet-based hardware and software.

The most important contribution of the paper lies in its use of networks to study empirically the best response demand curve for research scientists and the supply of research output by nodes of the networks. Using the linear model of social interactions, in which centrality of nodes plays a critical role, we find that

more central nodes employ more research scientists and, through the employment decision, produce more output. These results rely on instruments comprised of randomly assigned centralities from other nodes in previous networks, which break the correlation between actual centralities and the error term in the demand for research scientists; and on instruments based on national trends in science funding, which increase research output and are not correlated with the error term in the supply of research output equation.

Since centrality increases with citations from other nodes, research in different nodes forms strategic complements. By supporting research, science policy creates a system of knowledge externalities.

This research could be extended in several ways. Within a field, if a few highly central nodes contribute to the rest of the citation network more than the rest, is a more or less concentrated distribution of funding a more effective science policy toward that field? The question is not that simple to answer, because centralities depend on the level of funding, requiring a primitive concept of centrality. Identifying nodes by their betweenness or previously existing prominence, besides their degree and PageRank centralities, might help to answer this question. A second question concerns whether the between field allocation of funding is ideal from the standpoint of economic growth. Should the level of funding be adjusted in favor of one or more fields and against others, to maximize economic growth?

A worldwide analysis would add numerous within- and between-country dimensions to the network analysis in this paper. It might detect, for example, the changing relative centralities of fields and countries over time. It could also uncover enduring differences in the importance of fields in different nations. Do the engineering disciplines vary, for instance, in their centralities in Europe and Asia compared to North America?

Another set of questions has to do with language and international relations and their bearing on science citations and collaborations. How stringent are language barriers in the flow of scientific information between countries? The barrier is significantly greater in the case of pictographic and alphabetic languages than it is between different members of the European family of languages. And how do political competition and economic rivalries affect these knowledge flows? And does a state of war

upend or at least restrict the transmission of basic scientific knowledge until better times arrive? This is a full agenda indeed for the social and economic network analysis of science.

References

Acemoglu, Daron, Camilo García-Jimeno, and James A. Robinson (2015). "State Capacity and Economic Development: A Network Approach." **American Economic Review** 105(8): 2364-2409.

Ackerberg, D. A., K. Caves, and G. Frazer. (2015). "Identification Properties of Recent Production Function Estimators." **Econometrica** 83(6): 2411-2451.

Adams, James D., Grant C. Black, J. Roger Clemmons, and Paula E. Stephan. (2005). "Scientific Teams and Institutional Collaboration: Evidence from U.S. Universities, 1981-1999." **Research Policy** 34 (3): 259-285.

_____. (2010). "Is the U.S. Losing Its Preeminence in Higher Education?" **American Universities in a Global Market**, edited by Charles T. Clotfelter, University of Chicago Press for NBER, pages 33-68.

American Men and Women of Science: A Biographical Dictionary of Today's Leaders, 31st Edition. (2013). Farmington Hills, MI: Gale Cengage.

Autor, David H., Frank Levy, and Richard J. Murnane. (2003). "The Skill Content of Recent Technological Change: An Empirical Exploration." **The Quarterly Journal of Economics** 118(4): 1279-1333.

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou (2006). "Who's Who in Networks? Wanted: The Key Player." **Econometrica** 74: 1403-1417.

_____ (2010). "Delinquent Networks." **Journal of the European Economic Association** 8 (1): 34-61.

Battaglini, Marco, and Eleanora Patacchini. (2018). "Influencing Connected Legislators." **Journal of Political Economy** 126: 2277-2322.

Becker, Gary S. (1985). "Human Capital, Effort, and the Sexual Division of Labor." **Journal of Labor Economics** 3(1, Supplement Part 2): S33-S58.

Belhaj, Mohamed Moez, Yann Bramoullé, and Frédéric Deroïan (2014). "Network games under strategic complementarities." **Games and Economic Behavior** 88: 310-319.

Blume, L.E., W. A. Brock, S. N. Durlauf, and R. Joyaraman. (2015). "Linear Social Interactions Models." **Journal of Political Economy** 123(2): 444-496.

Bonacich, Phillip (1987). "Power and Centrality: A Family of Measures." **American Journal of Sociology** 92: 1170-1182.

Borgatti, Stephen P. (2005). "Centrality and Network Flow." **Social Networks** 27: 55-71.

_____, and Martin G. Everett (2006). "A Graph-Theoretic Perspective on Centrality." **Social Networks** 28: 466-484.

Borjas, George (1987). "Self-Selection and the Earnings of Immigrants." **American Economic Review** 77(4): 531-553.

Boucher, Vincent, Michelle Rendall, Philip Ushchev, and Yves Zenou (2024). "Toward a General Theory of Peer Effects." **Econometrica** 92: 543-565.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009). "Identification of Peer Effects Through Social Networks." **Journal of Econometrics** 150: 41-55.

_____, Rachel Kranton, and Martin D'Amours (2014). "Strategic Interaction and Networks." **American Economic Review** 104(3): 898-930.

Brandes, Ulrik (2008). "On Variants of Shortest-Path Betweenness Centrality and Their Generic Computation." **Social Networks** 30: 136-145.

Breschi, Stefano, and Francesco Lissoni. (2009). "Mobility of Skilled Workers and Co-Invention Networks: An Anatomy of Localised Knowledge Flows." **Journal of Economic Geography** 9(4): 439-468.

Brin, Sergey, and Lawrence Page. (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine." **Computer Networks** 36: 107-117.

Bulow, Jeremy I., John D. Geanakoplos, and Paul D. Klemperer. (1985). "Multimarket Oligopoly: Strategic Substitutes and Complements." **Journal of Political Economy** 95: 488-511.

Caballero, Ricardo J., Eduardo M.R.A. Engel, and John Haltiwanger (1997). "Aggregate Employment Dynamics: Building from Microeconomic Evidence," **American Economic Review** 87(1): 115-137.

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou (2009). "Peer Effects and Social Networks in Education." **Review of Economic Studies** 76: 1239-1267.

Carvalho, Vasco M., and Alireza Tahbaz-Salehi (2019). "Production Networks: A Primer." **Annual Reviews of Economics** 11: 635-63.

Clauset, A., M.E.J. Newman, and C. Moore. (2004). "Finding Community Structure in Very Large Networks," **Physical Review E** 79: 066111.

Cormen, Thomas H.; Charles E. Leiserson, Ronald E. Rivest, and Clifford Stein. (2001). "Section 24.3: Dijkstra's algorithm." **Introduction to Algorithms (2nd ed.)**. Cambridge, Mass.: MIT Press and McGraw–Hill.

"Dijkstra's Algorithm." **Wikipedia, The Free Encyclopedia**. Wikimedia Foundation, https://en.wikipedia.org/wiki/Dijkstra%27s_algorithm, Accessed November 17, 2019.

Elliott, Matthew, and Benjamin Golub. (2022). "Networks and Economic Fragility." **Annual Review of Economics** 14: 665-96.

Fainmesser, Itay P., and Andrea Galeotti. (2020). "Pricing Network Effects: Competition." **American Economic Journal: Microeconomics** 12 (3): 1-32.

Freeman, Linton C. (1977), "A Set of Measures of Centrality Based on Betweenness." **Sociometry** 40: 35-41.

Goyal, Sanjeev, and José Luis Moraga-González (2001). "R&D Networks." **RAND Journal of Economics** 32 (4): 686-707.

Griliches, Zvi. (1979). “Issues in Addressing the Contribution of Research and Development to Productivity Growth.” **Bell Journal of Economics** 10 (1): 92-116.

_____ (1992). “The Search for R&D Spillovers.” **Scandinavian Journal of Economics** 94 (Supplement): S29-S47.

Gross, Daniel P., and Bhaven N. Sampat. (2023). “America, Jump-Started: World War II R&D and the Takeoff of the U.S. Innovation System.” Working Paper, Duke University and Columbia University.

Hamermesh, Daniel. (1989). “Labor Demand and the Structure of Adjustment Costs.” **American Economic Review** 79(4): 674-689.

Immorlica, Nicole, Rachel Kranton, Mihai Manea, and Greg Stoddard. (2017). “Social Status in Networks.” **American Economic Journal: Microeconomics** 9 (1): 1-30.

Jackson, Matthew O., and Asher Wolinsky (1996). “A Strategic Model of Social and Economic Networks.” **Journal of Economic Theory** 71: 44-74.

_____ (2008). **Social and Economic Networks**. Princeton, NJ: Princeton University Press.

_____, Brian W. Rogers, and Yves Zenou (2017). “The Economic Consequences of Social-Network Structure.” **Journal of Economic Literature** 55 (1): 49-95.

Jaffe, Adam B., and Manuel Trajtenberg. (2002). **Patents, Citations and Innovations: A Window on the Knowledge Economy**. Cambridge, Massachusetts: Massachusetts Institute of Technology Press.

Katz, Leo. (1953). “A New Status Index Derived from Sociometric Analysis.” **Psychometrica** 18: 39-43.

Köenig, Michael D., Xiaodong Liu, and Yves Zenou. (2019). “R&D Networks: Theory, Empirics, and Policy Implications.” **The Review of Economics and Statistics** 101: 476-491.

Langlois, Richard N., and David C. Mowery. (1996). "The Federal Government Role in the Development of the U.S. Software Industry." **The International Software Industry,** edited by David C. Mowery, New York: Oxford University Press, 53-85.

Lécuyer, Christophe. (2014). "Semiconductor Innovation and Entrepreneurship at Three University of California Campuses." In **Public Universities and Regional Growth: Insights from the University of California**, edited by Martin Kenney and David C. Mowery. Stanford, California: Stanford Business Books, Stanford University Press.

Levinsohn, J., and A. Petrin. (2003). "Estimating Production Functions Using Inputs to Control for Unobservables." **Review of Economic Studies** 70: 317-341.

Li, Guan-Cheng (2013). "Disambiguation of Inventors, USPTO, 1975-2013," Berkeley, California: Fung Institute of Engineering Leadership, UC Berkeley College of Engineering, September 17, 2013.

Manski, Charles F. (1993). "Identification of Endogenous Social Effects: The Reflection Problem." **Review of Economic Studies** 60: 419-431.

Mowery, David C., and Nathan Rosenberg. (1998). **Paths of Innovation: Technological Change in 20th Century America**. New York: Cambridge University Press.

_____, Richard R. Nelson, Bhaven N. Sampat, and Arvids A. Ziedonis. (2004). **Ivory Tower and Industrial Innovation: University-Industry Technology Transfer Before and After the Bayh-Dole Act.** Stanford, California: Stanford Business Books, Stanford University Press.

National Research Council. (2011). **A Data-Based Assessment of Research-Doctorate Programs in the United States**. Washington, DC" National Academies Press.

National Science Board. **Science and Engineering Indicators** (Various Years). Washington, DC: National Science Foundation.

Newman, M.E.J. and M. Girvan (2004). "Finding and Evaluating Community Structure in Networks." **Phys. Rev. E** 69, 026113 (2004).

_____. (2005). "A Measure of Betweenness Centrality Based on Random Walks." **Social Networks** 27: 39-54.

_____ (2010). **Networks: An Introduction**. Oxford, UK: Oxford University Press.

Olley, S., and A. Pakes. (1996). "The Dynamics of Productivity in the Telecommunications Equipment Industry." **Econometrica** 64: 1263-1298.

"PageRank," **Wikipedia, The Free Encyclopedia**, Wikimedia Foundation, https://en.wikipedia.org/wiki/PageRank, Accessed November 17, 2019.

Roy, A.D. (1951). "Some Thoughts on the Distribution of Earnings." **Oxford Economic Papers (New Series)** 3: 135-146.

Wasserman, Stanley, and Katherine Faust. (1994). **Social Network Analysis: Methods and Applications**. New York: Cambridge University Press.

Web of Science ™ Database (Various Years). Philadelphia, Pennsylvania: Clarivate Analytics.

Table 1
Citation Networks: Density and Number of Components,
Initial and Final Periods Compared

Science Group or Field	Unweighted Network Density [a]		Number of Components [b]	
	1981-1986	2006-2011	1981-1986	2006-2011
Life Sciences				
Agriculture	0.531	0.722	1	1
Biology	0.409	0.569	7	4
Medicine	0.364	0.506	6	3
Mathematical Sciences				
Computer Science	0.124	0.361	6	1
Mathematics and Statistics	0.197	0.407	1	2
Physical Sciences				
Astronomy	0.619	0.836	2	1
Chemistry	0.252	0.442	6	2
Physics	0.333	0.522	7	3
Geological Sciences				
Atmospheric Science	0.472	0.913	1	1
Earth Science	0.358	0.707	2	1
Oceanography	0.517	0.802	2	1
Engineering				
Aeronautical	0.101	0.227	6	1
Chemical	0.063	0.106	9	3
Civil	0.034	0.237	10	1
Electrical	0.082	0.266	8	2
Materials Science	0.146	0.392	5	3
Mechanical	0.028	0.088	8	5
Other	0.075	0.318	11	1
Social and Behavioral Sciences				
Economics	0.227	0.573	1	1
Psychology	0.484	0.809	1	1
All Fields	0.046	0.118	2	1

Notes. Data are derived from Clarivate Analytics (formerly Thomson-Reuters) Web of Science. Data consists of six successive citation networks ordered by field of science, for six-year intervals from 1981-1986, 1986-1991, …, 2006-2011. Fields correspond to National Science Foundation classifications. The table reports statistics for 1981-1986 and 2006-2011, enabling comparisons over a 25-year period. [a] Density is the actual number of links between nodes divided by the maximum possible number of links. Throughout this paper, nodes are institutions and fields meeting a size cutoff. [b] A component is a set of nodes in a network that are connected to each other but not to other nodes in the network.

Table 2
Flow of Citations Within and Between Fields:
Initial and Final Periods Compared

Science Group or Field	1981-1986				2006-2011			
	Within Field	From Other Fields	To Other Fields	Net Inflow	Within Field	From Other Fields	To Other Fields	Net Inflow
Life Sciences								
Agriculture	21.8	14.2	16.2	-	59.4	54.9	67.3	-
Biology	304.9	159.7	131.7	+	1,436.9	762.0	627.6	+
Medicine	510.5	130.4	153.2	-	2,830.8	616.0	669.1	-
Mathematical Sciences								
Computer Science	4.8	4.2	4.4	-	14.5	20.7	20.4	+
Mathematics and Statistics	7.5	5.3	4.4	+	30.0	27.3	22.6	+
Physical Sciences								
Astronomy	25.7	6.9	7.0	-	271.4	28.0	28.5	-
Chemistry	70.2	38.7	45.6	-	349.0	196.5	244.7	-
Physics	133.6	40.5	34.9	+	598.0	162.8	153.9	+
Geological Sciences								
Atmospheric Science	7.1	4.4	3.4	+	61.7	23.8	19.2	+
Earth Science	23.2	12.3	14.1	-	122.0	60.0	68.9	-
Oceanography	1.5	2.2	1.8	+	3.7	7.8	6.4	+
Engineering								
Aeronautical	0.3	0.5	0.5	0	0.9	1.7	1.7	0
Chemical	0.9	2.0	2.0	0	2.0	6.6	8.5	-
Civil	0.3	1.1	1.2	-	4.2	5.6	7.0	-
Electrical	4.1	7.2	7.3	-	25.2	25.5	31.6	-
Materials Science	2.7	3.7	3.9	-	17.7	33.0	44.0	-
Mechanical	0.5	0.9	1.8	-	1.2	5.6	6.3	-
Other	2.0	4.3	5.3	-	13.1	32.7	42.2	-
Social and Behavioral Sciences								
Economics	5.2	2.7	2.1	+	23.0	17.5	12.0	+
Psychology	28.4	16.0	16.4	-	157.5	102.0	108.4	-
All Fields	1,154.9	457.3	457.3	0	6,022.4	2,190.2	2,190.2	0

Notes. Data are derived from Clarivate Analytics (formerly Thomson-Reuters) Web of Science. Statistics are total citation flows, in thousands rounded to the nearest hundred. Data consist of six successive networks, in which citations link pairs of nodes, for six-year intervals from 1981-1986, 1986-1991, …, 2006-2011. Table reports statistics for 1981-1986 and 2006-2011, enabling comparisons 25 years apart.

Table 3
Descriptive Statistics
Panel of Institution-Location-Field Nodes [a], 1981-2011

Variables	Obs.	Mean	Std. Dev.	Min	Max
Best Response Demand for Research Scientists [b]					
All Research scientists	77,048	260.60	809.13	0	36,973
Long-Term Research scientists	77,048	179.61	548.86	0	21,861
Citation In-Degree Centrality [c]	59,862	1,687.54	5,929.13	0	239,280
Citation PageRank Centrality [d]	59,862	0.01032	0.01537	0.00066	0.20935
Other Fields (1 if Yes, 0 if No)	77,048	0.985	0.120	0	1
Other Locations (1 if Yes, 0 if No)	77,048	0.183	0.387	0	1
Defense Exposure × Defense R&D Cycle:					
Defense Intramural Dummy × Intramural R&D (1981=1.0)	77,048	0.102	0.380	0	2.184
Defense Contractor Dummy × Contractor R&D (1981=1.0)	77,048	0.012	0.106	0	1.457
Average Distance to Top Research Cities	77,048	1,163.03	320.19	852.24	1,847.02
Research Scientists in Same Field elsewhere in the Same MSA	77,048	307.77	1,182.19	0	30,834.66
Research Output Equation [b]					
Citations Received, Next 5 Years	64,447	490.55	1,808.16	0	80,018.61
All Research scientists Squared	64,447	575,554.7	1.03e+07	0	1.09e+09
Long-Term Research scientists Squared	64,447	332,459	5,842,769	0	4.78e+08

Notes. The time period is 1981-2011 for Best Response Demand equation, but it is 1981-2006 for the Research Output equation, since citations received over the next five years are not available for 2007-2011. [a] University nodes are institution-fields that publish at least 60 papers per period. Non-university nodes are institution-location-fields whose institution-fields publish at least 400 papers in a period, and whose institution-location-fields publish at least 100 papers in a period. Institutions are universities, firms, government agencies and laboratories, and private nonprofit biomedical research centers. Data are derived from the Web of Science ™ of Clarivate Analytics (formerly Thomson-Reuters). Citation networks are arranged over periods from 1981-1986, 1986-1991, …, 2006-2011. [b] Regression equations include dummies for institution, field, and year, besides variables shown in the table. [c] Citation In-Degree Centrality is the sum of citations received by a department over a time interval of six years. [d] Citation PageRank Centrality sums to 1.0 over departments owing to the normalization imposed by the PageRank definition.

Table 4
Centrality Instruments

Instrument Name	Description	Explanation
Centrality IV, Same Field, Lag=1	Randomly assigned centrality for the same field and institution type, from previous period's network.	Set random seed to initialize random sample from the uniform distribution. For the population, we use other institutions in the same field and type of institution in the previous period. [a]
Centrality IV, Other Field, Lag=1	Randomly assigned centrality from a different field that is most cited by the focal field, for same institution type, from previous period's network.	Set random seed to initialize random sample from the uniform distribution. For the population, use other institutions in the field citing this field most, using the previous period. [b]
Centrality IV, Same Field, Lag=2	Randomly assigned centrality for the same field and institution type, from the network two periods before.	Set random seed to initialize random sample from the uniform distribution. For the population, use other institutions in the same field from two periods before.
Centrality IV, Other Field, Lag=2	Randomly assigned centrality from a different field that is most cited by the focal field, for same institution type, from the network two periods before.	Set random seed to initialize random sample from the uniform distribution. For the population, use other institutions in the field citing this field most, using network from two periods before.

Notes. Instruments for the centralities are taken from the citation networks of research departments previously described, for periods defined by 1981-1986, 1986-1991, …, and 2006-2011. Citation networks do not overlap, because 1986 papers in the interval 1981-1986 are not cited until the interval 1986-1991, and because citations are within periods. The institution types are Top 200 universities, Top 100 R&D-intensive firms, federal agencies and federal laboratories, and biomedical research centers. [a] Sampling for medicine in university A during 1986-1991 selects the centrality in a different university F, in medicine, for the period 1981-1986. [b] Because the most cited field for medicine is biology, sampling for medicine in university A during 1986-1991 selects the centrality in a different university R, in biology, for the period 1981-1986. [c] Sampling for medicine in university A during 1991-1996 selects the centrality in a different university K, in medicine, for the period 1981-1986. [d] Because the most cited field for medicine is biology, sampling for medicine in university A during 1991-1996 selects the centrality in a different university Z, in biology, for the period 1981-1986. Notice that the one period lag eliminates the period 1981-1986, while the two-period lag eliminates both 1981-1986 and 1986-1991.

Table 5
Best Response Demand for Research Scientists
Citation In-Degree Centrality
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(5.1)	(5.2)	(5.3)	(5.4)	(5.5)	(5.6)
Dependent Variable	All Research Scientists			Long-Term Research Scientists		
Time Period Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Science Field Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Institution Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Method	OLS	2SLS	2SLS	OLS	2SLS	2SLS
VCE Assumption	Robust	Robust	Clustered [a]	Robust	Robust	Clustered [a]
Lag on Instruments for Citation In-Degree Centrality	--	1 Period	1 Period	--	1 Period	1 Period
Citation In-Degree Centrality	0.137**	0.148**	0.148**	0.085**	0.079**	0.079**
	(0.002)	(0.006)	(0.011)	(0.003)	(0.006)	(0.010)
Same Institution: Other Fields, Same Location (1 if Yes, 0 if No)	259.59**	209.49**	209.49**	203.59**	231.15**	231.15**
	(17.59)	(31.21)	(70.60)	(17.66)	(32.48)	(65.71)
Same Institution: Other Locations, Same Field (1 if Yes, 0 if No)	-44.34**	-25.87*	-25.87	-45.27**	-55.43**	-55.43*
	(5.95)	(11.26)	(26.38)	(6.12)	(11.72)	(23.09)
Defense Intramural Dummy × Intramural R&D (1981=1.0)	-4.79	3.70	3.70	68.41**	63.73**	63.73**
	(14.42)	(13.97)	(19.09)	(11.89)	(12.93)	(19.66)
Defense Contractor Dummy × Contractor R&D (1981=1.0)	4.37	-11.76	-11.76	32.18**	41.05**	41.05
	(12.82)	(14.65)	(25.58)	(12.03)	(15.39)	(21.36)
Average Distance, Top Research Cities	-0.052**	-0.042**	-0.042	-0.051**	-0.057**	-0.057
	(0.009)	(0.011)	(0.034)	(0.008)	(0.010)	(0.033)
Research Scientists in Same Field and Same MSA	-0.039**	-0.044**	-0.044**	-0.029**	-0.026**	-0.026**
	(0.003)	(0.005)	(0.012)	(0.003)	(0.004)	(0.008)
F-Test for Excluded Instruments		32.374	8.649		32.374	8.659
[P-Value]		[0.0000]	[0.0002]		[0.0000]	[0.0002]
Test of Overidentifying Restrictions		0.436			3.430	
[P-Value]		[0.509]			[0.0640]	
Number of Observations	51,117	51,117	51,117	51,117	51,117	51,117
F (284, 50,892)	495.51			340.71		
Probability>F	0.0000			0.0000		
Wald Chi-Square		134,793.23	19,477.53		96,621.68	1,118,846.82
[P-Value]		[0.0000]	[0.0000]		[0.0000]	[0.0000]

Table 5
Best Response Demand for Research Scientists
Citation In-Degree Centrality
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(5.1)	(5.2)	(5.3)	(5.4)	(5.5)	(5.6)
R-squared	0.927	0.924	0.924	0.844	0.842	0.842
Root MSE	261.96	266.52	266.52	259.04	259.95	259.95

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. The period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates. [a] Clustering is by institution and field.
**Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level.

Table 6
Best Response Demand for Research Scientists
Interactions Between Institutional Type and Citation In-Degree Centrality
2SLS Robust Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(6.1)	(6.2)	(6.3)	(6.4)
Dependent Variable	All Research Scientists			
Institutional Type	Defense Institutions	Government Institutions	Firms	Biomedical Centers
Panel A. Comparison Group: Rest of the sample				
Citation In-Degree Centrality	0.148**	0.147**	0.146**	0.151**
	(0.006)	(0.006)	(0.006)	(0.006)
Citation In-Degree Centrality ×	0.023	0.002	-0.094*	-0.047**
Dummy for Institutional Type	(0.013)	(0.008)	(0.043)	(0.006)
Tests for Weak Instruments:				
Minimum Eigenvalue Statistic	112.031	114.842	38.845	106.549
Critical Values, Weak Instruments:				
2SLS Relative Bias of 5%	11.04	11.04	11.04	11.04
2SLS Size of 10% Wald Test	16.87	16.87	16.87	16.87
Number of Observations	51,117	51,117	51,117	51,117
Panel B. Comparison Group: Universities				
Citation In-Degree Centrality	0.138**	0.147**	0.125**	0.096**
	(0.012)	(0.008)	(0.012)	(0.023)
Citation In-Degree Centrality ×	0.015	0.002	-0.204**	-0.010
Dummy for Institutional Type	(0.017)	(0.008)	(0.075)	(0.017)
Test for Weak Instruments:				
Minimum Eigenvalue Statistic	26.33	61.391	15.189	12.719
Critical Values, Weak Instruments:				
2SLS Relative Bias of 5%	11.04	11.04	11.04	11.04
2SLS Size of 10% Wald Test	16.87	16.87	16.87	16.87
Number of Observations	43,109	46,601	43,619	40,733

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. Periods 1981-1986 and 1986-1991 are deleted since IVs are lagged two periods. See the text and Table 3 for a description of the covariates. **Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level. Both the minimum eigenvalue statistic and the critical values for weak instruments are F-statistics. If the minimum eigenvalue exceeds the critical values, the hypothesis of weak instruments is rejected.

Table 7
Best Response Demand for Research Scientists
Citation In-Degree Centrality
Panel IV, Fixed Effects Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(7.1)	(7.2)
Dependent Variable	All Research Scientists	Long-Term Research Scientists
Time Period Fixed Effects	Yes	Yes
Science Field Fixed Effects	Yes	Yes
Institution Fixed Effects	Yes	Yes
Method	Panel IV, FE [a]	Panel IV, FE [a]
VCE Assumption	Clustered [b]	Clustered [b]
Lag on Instruments for Citation In-Degree Centrality	1 Period	1 Period
Citation In-Degree Centrality	0.105**	0.029
	(0.009)	(0.016)
Same Institution: Other Fields, Same Location (1 if Yes, 0 if No)	48.96	37.28**
	(28.49)	(14.14)
Same Institution: Other Locations, Same Field (1 if Yes, 0 if No)	39.54	58.23*
	(21.59)	(26.55)
Defense Intramural Dummy × Intramural R&D (1981=1.0)	-26.40	28.59
	(13.70)	(17.76)
Defense Contractor Dummy × Contractor R&D (1981=1.0)	47.45**	98.34**
	(15.31)	(18.81)
Average Distance, Top Research Cities	--	--
Research Scientists in Same Field and Same MSA	0.000	0.019
	(0.014)	(0.020)
Number of Observations	51,117	51,117
Number of Groups	2,230	2,230
Wald Chi-Square	1,630.58	13,121.07
[P-Value]	0.0000	0.0000
Within R-Squared	0.825	0.825
Between R-Squared	0.917	0.917

Table 7
Best Response Demand for Research Scientists
Citation In-Degree Centrality
Panel IV, Fixed Effects Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(7.1)	(7.2)
Sigma u (Within Root MSE)	361.10	361.10
Sigma e (Between Root MSE)	133.75	133.75
Rho (Fraction of Variance Due to Within Variance)	0.879	0.879

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. The period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates. [a] Panel IV is the within or fixed effects regression where the centrality is treated as endogenous. [b] Clustering is by institution and field. **Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level.

Table 8
Best Response Demand for Research Scientists
Citation PageRank Centrality
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(8.1)	(8.2)	(8.3)	(8.4)	(8.5)	(8.6)
Dependent Variable	All Research Scientists			Long-Term Research Scientists		
Time Period Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Science Field Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Institution Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Method	OLS	2SLS	2SLS	OLS	2SLS	2SLS
VCE Assumption	Robust	Robust	Clustered [a]	Robust	Robust	Clustered [a]
Lag on Instruments for Citation PageRank Centrality	1 Period	1 Period	1 Period	1 Period	1 Period	1 Period
Citation PageRank Centrality	16,418.93**	28,001.69**	28,001.69*	11,336.69**	17,191.56**	17,191.56*
	(1,024.31)	(4,787.45)	(13,063.00)	(677.22)	(3,167.32)	(8,275.20)
Same Institution: Other Fields, Same Location (1 if Yes, 0 if No)	910.47**	898.72**	898.72**	604.23**	598.29**	598.29**
	(49.89)	(49.49)	(186.17)	(32.20)	(31.99)	(124.13)
Same Institution: Other Locations, Same Field (1 if Yes, 0 if No)	-321.85**	-344.00**	-344.00**	-218.90**	-230.09**	-230.09**
	(15.30)	(17.82)	(69.60)	(10.27)	(11.89)	(46.81)
Defense Intramural Dummy × Intramural R&D (1981=1.0)	-118.99**	-119.73**	-119.73**	-2.18	-2.55	-2.55
	(35.16)	(37.16)	(29.05)	(24.08)	(24.95)	(19.75)
Defense Contractor Dummy × Contractor R&D (1981=1.0)	182.83**	157.17**	157.17**	139.71**	126.74**	126.74**
	(35.08)	(33.56)	(60.77)	(25.28)	(24.54)	(39.88)
Average Distance, Top Research Cities	-0.100**	-0.030	-0.030	-0.073**	-0.038	-0.038
	(0.028)	(0.041)	(0.146)	(0.019)	(0.027)	(0.095)
Research Scientists in Same Field and Same MSA	0.036**	0.034**	0.034	0.016**	0.015**	0.015
	(0.008)	(0.008)	(0.030)	(0.005)	(0.005)	(0.019)
F-Test for Excluded Instruments		25.686	3.429		25.686	3.429
[P-Value]		[0.0000]	[0.0326]		[0.0000]	[0.0326]
Test of Overidentifying Restrictions		0.063			0.055	
[P-Value]		[0.8022]			[0.8149]	
Number of Observations	51,177	51,117	51,117	51,177	51,117	51,117
F (285,50891)	151.25			183.27		
[P-Value]	[0.0000]			[0.0000]		

Table 8
Best Response Demand for Research Scientists
Citation PageRank Centrality
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(8.1)	(8.2)	(8.3)	(8.4)	(8.5)	(8.6)
Wald Chi-Square		43,003.83	65,246.54		50,618.41	7.88e+07
[P-Value]		[0.0000]	[0.0000]		[0.0000]	[0.0000]
R-squared	0.423	0.409	0.409	0.431	0.423	0.423
Root MSE	735.41	742.19	742.19	494.55	496.53	496.53

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. The period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates. [a] Clustering is by institution and field. **Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level.

Table 9
Best Response Demand for Research Scientists
Citation In-Degree Centrality
2SLS Estimates, Two Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(9.1)	(9.2)	(9.3)	(9.4)
Dependent Variable	All Research Scientists		Long-Term Research Scientists	
Time Period Fixed Effects	Yes	Yes	Yes	Yes
Science Field Fixed Effects	Yes	Yes	Yes	Yes
Institution Fixed Effects	Yes	Yes	Yes	Yes
Method	2SLS	2SLS	2SLS	2SLS
VCE Assumption	Robust	Clustered [a]	Robust	Clustered [a]
Lag on Instruments for Citation In-Degree Centrality	2 Periods	2 Periods	2 Periods	2 Periods
Citation In-Degree Centrality	0.138**	0.138**	0.080**	0.080**
	(0.005)	(0.010)	(0.005)	(0.009)
Same Institution: Other Fields, Same Location (1 if Yes, 0 if No)	259.97**	259.97**	224.20**	224.20**
	(31.22)	(79.27)	(34.10)	(72.15)
Same Institution: Other Locations, Same Field (1 if Yes, 0 if No)	-36.60**	-36.60	-50.45**	-50.45
	(11.63)	(29.79)	(12.74)	(26.76)
Defense Intramural Dummy × Intramural R&D (1981=1.0)	-7.93	-7.93	67.53**	67.53**
	(15.46)	(20.14)	(13.59)	(20.56)
Defense Contractor Dummy × Contractor R&D (1981=1.0)	-1.77	-1.77	53.48*	53.48*
	(21.28)	(21.90)	(24.51)	(24.12)
Average Distance, Top Research Cities	-0.057**	-0.057	-0.060**	-0.060
	(0.012)	(0.039)	(0.011)	(0.033)
Research Scientists in Same Field and Same MSA	-0.039**	-0.039**	-0.027**	-0.027**
	(0.004)	(0.012)	(0.004)	(0.009)
F-Test for Excluded Instruments	38.26	6.108	38.26	6.108
[P-Value]	[0.0000]	[0.0023]	[0.0000]	[0.0023]
Test of Overidentifying Restrictions	0.783		4.166	
[P-Value]	[0.3762]		[0.0413]	
Number of Observations	42,150	42,150	42,150	42,150

Table 9
Best Response Demand for Research Scientists
Citation In-Degree Centrality
2SLS Estimates, Two Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(9.1)	(9.2)	(9.3)	(9.4)
Wald Chi-Square	151,083.18	364,036.46	84,816.29	774,951.30
[P-Value]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
R-squared	0.930	0.930	0.844	0.844
Root MSE	273.65	272.83	272.47	272.47

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. Periods 1981-1986 and 1986-1991 are deleted since IVs are lagged by two periods. See the text and Table 3 for a description of the covariates. [a] Clustering is by institution and field. **Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level.

Table 10
Best Response Demand for Research Scientists
Citation PageRank Centrality
2SLS Estimates, Two Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(10.1)	(10.2)	(10.3)	(10.4)
Dependent Variable	All Research Scientists		Long-Term Research Scientists	
Time Period Fixed Effects	Yes	Yes	Yes	Yes
Science Field Fixed Effects	Yes	Yes	Yes	Yes
Institution Fixed Effects	Yes	Yes	Yes	Yes
Method	2SLS	2SLS	2SLS	2SLS
VCE Assumption	Robust	Clustered [b]	Robust	Clustered [b]
Lag on Instruments for PageRank Citation Centrality	2 Periods	2 Periods	2 Periods	2 Periods
Citation PageRank Centrality	37,754.43**	37,754.43*	22,831.9**	22,831.9*
	(6,871.3)	(17,966.7)	(4,290.32)	(11,164.07)
Institution: Other Fields, Same Location (1 if Yes, 0 if No)	993.97**	993.97**	647.92**	647.92**
	(58.61)	(209.82)	(37.44)	(136.85)
Institution: Other Locations, Same Field (1 if Yes, 0 if No)	-380.63**	-380.63**	-250.69**	-250.69**
	(20.25)	(74.68)	(13.16)	(48.82)
Defense Intramural Dummy × Intramural R&D (1981=1.0)	-137.71**	-137.71**	-6.83	-6.83
	(43.29)	(39.43)	(28.20)	(24.64)
Defense Contractor Dummy × Contractor R&D (1981=1.0)	128.27	128.27*	127.59*	127.59**
	(80.24)	(57.23)	(55.75)	(32.65)
Average Distance, Top Research Cities	0.011	0.011	-0.015	-0.015
	(0.055)	(0.177)	(0.035)	(0.113)
Research Scientists in Same Field and Same MSA	0.025**	0.025	0.010*	0.010
	(0.008)	(0.030)	(0.005)	(0.019)
F-Test for Excluded Instruments	19.48	3.39	19.48	3.39
[P-Value]	[0.0000]	[0.0339]	[0.0000]	[0.0339]
Test of Overidentifying Restrictions	2.749		1.597	
[P-Value]	[0.0973]		[0.2063]	
Number of Observations	42,150	42,150	42,150	42,150
Wald Chi-Square	52,698.11	9,432.81	49,037.08	2.03e+09
[P-Value]	[0.0000]	[0.0000]	[0.0000]	[0.0000]

Table 10
Best Response Demand for Research Scientists
Citation PageRank Centrality
2SLS Estimates, Two Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(10.1)	(10.2)	(10.3)	(10.4)
R-squared	0.399	0.399	0.414	0.414
Root MSE	799.71	799.71	528.8	528.8

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. Periods 1981-1986 and 1986-1991 are deleted since IVs are lagged two periods. See the text and Table 3 for a description of the covariates. [a] Panel IV is the within or fixed effects regression where the centrality is treated as endogenous. [b] Clustering is by institution and field. **Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level.

Table 11A
Two Equation Systems:
Best Response Demand for Research Scientists and Research Output Equation
Citation In-Degree Centrality
3SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Descriptor	System 1 All Research Scientists		System 2 Long-Term Research Scientists	
Panel A. University-Field-Year Funding Instruments				
Dependent Variable	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]
Method	3SLS		3SLS	
VCE Assumption	Clustered		Clustered	
Lag on Instruments for Citation Centrality	1 Period		1 Period	
Funding Instruments for Research Scientists	University-Field-Year Level		University-Field-Year Level	
Citation In-Degree Centrality	0.1429**			
	(0.0055)			
All Research Scientists Squared, Lagged One Year		0.000282**		
		(0.0000815)		
Citation In-Degree Centrality			0.1090**	
			(0.0044)	
Long-Term Research Scientists Squared, Lagged One Year				0.000485**
				(0.000143)
Number of Observations	36,086	36,086	36,086	36,086
Chi-Square	1.36e+08	6275	9.15e+06	6408
[P-Value]	0.0000	0.0000	0.0000	0.0000

Table 11A
Two Equation Systems:
Best Response Demand for Research Scientists and Research Output Equation
Citation In-Degree Centrality
3SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Descriptor	System 1 All Research Scientists		System 2 Long-Term Research Scientists	
Panel B. Field-Year Funding Instruments				
Dependent Variable	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]
Method	3SLS		3SLS	
VCE Assumption	Clustered [c]		Clustered [c]	
Lag on Instruments for Citation Centrality	1 Period		1 Period	
Instruments for Research Scientists	Field-Year Level		Field-Year Level	
Citation In-Degree Centrality	0.1354** (0.0041)			
All Research Scientists Squared, Lagged One Year		0.000260** (0.0000951)		
Citation In-Degree Centrality			0.0924** (0.0032)	
Long-Term Research Scientists Squared, Lagged One Year				0.000483** (0.000190)
Number of Observations	37,684	37,684	37,684	37,684
Chi-Square	117,372.34	2871.78	547,805.95	6518.79
[P-Value]	0.0000	0.0000	0.0000	0.0000

Notes. Data are based on a panel of nodes, which are defined as institutions, locations, and fields. Data covers the years from 1986-2011. Period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates.

[a] Dependent variable in Eq. 1 is All Research Scientists (System 1) or Long-Term Research Scientists (System 2). [b] In both System 1 and System 2 the dependent variable in Eq. 2 is total citations received over the next five years. Equations follow the specification of Table 3. Both equations include fixed effects for institution, field, and year. [c] Clustering is by institution, location, and field. **Coefficient differs from zero at the one percent level and above. *Coefficient differs from zero at the five percent level.

Table 11B
Two Equation Systems:
Best Response Demand for Research Scientists, Research Output Equation
Citation Page Rank Centrality
3SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Descriptor	System 1 All Research Scientists		System 2 Long-Term Research Scientists	
Panel C. University-Field-Year Funding Instruments				
Dependent Variable	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]
Method	3SLS		3SLS	
VCE Assumption	Clustered		Clustered	
Lag on Instruments for Citation Centrality	1 Period		1 Period	
Funding Instruments for Research Scientists	University-Field-Year Level		University-Field-Year Level	
PageRank Citation Centrality	250,489**			
	(49,224.2)			
All Research Scientists Squared, Lagged One Year		0.000285**		
		(0.000081)		
Citation PageRank Centrality			191,140.5**	
			(37,763.87)	
Long-Term Research Scientists Squared, Lagged One Year				0.000490**
				(0.000141)
Number of Observations	36,086	36,086	36,086	36,086
Chi-Square	8.59e+06	7585.50	5.68e+07	5525.41
[P-Value]	0.0000	0.0000	0.0000	0.0000

Table 11B
Two Equation Systems:
Best Response Demand for Research Scientists, Research Output Equation
Citation Page Rank Centrality
3SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Descriptor	System 1 All Research Scientists		System 2 Long-Term Research Scientists	
Panel D. Field-Year Funding Instruments				
Dependent Variable	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]	Eq.1: Best Response Demand for Scientists [a]	Eq. 2: Research Output Equation, Citations Received [b]
Method	3SLS		3SLS	
VCE Assumption	Clustered		Clustered	
Lag on Instruments for Citation Centrality	1 Period		1 Period	
Instruments for Research Scientists	Field-Year Level		Field-Year Level	
Citation PageRank Centrality	100,084.5 (62,937.18)			
All Research Scientists Squared, Lagged One Year		0.000300** (0.000098)		
PageRank Citation Centrality			69,128.69 (43,641.33)	
Long-Term Research Scientists Squared, Lagged One Year				0.000574** (0.000199)
Number of Observations	37,684	37,684	37,684	37,684
Chi-Square	11,312.55	3722.50	15,671.96	2368.06
[P-Value]	0.0000	0.0000	0.0000	0.0000

Notes. Data are based on a panel of nodes, which are defined as institutions, locations, and fields. Data covers the years from 1986-2011. Period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates.

[a] Dependent variable in Eq. 1 is All Research Scientists (System 1) or Long-Term Research Scientists (System 2). [b] In both System 1 and System 2 the dependent variable in Eq. 2 is total citations received over the next five years. Equations follow the specification of Table 3. Both equations include fixed effects for institution, field, and year. [c] Clustering is by institution, location, and field. **Coefficient differs from zero at the one percent level and above. *Coefficient differs from zero at the five percent level.

Figure 1
Citation Densities Over Time

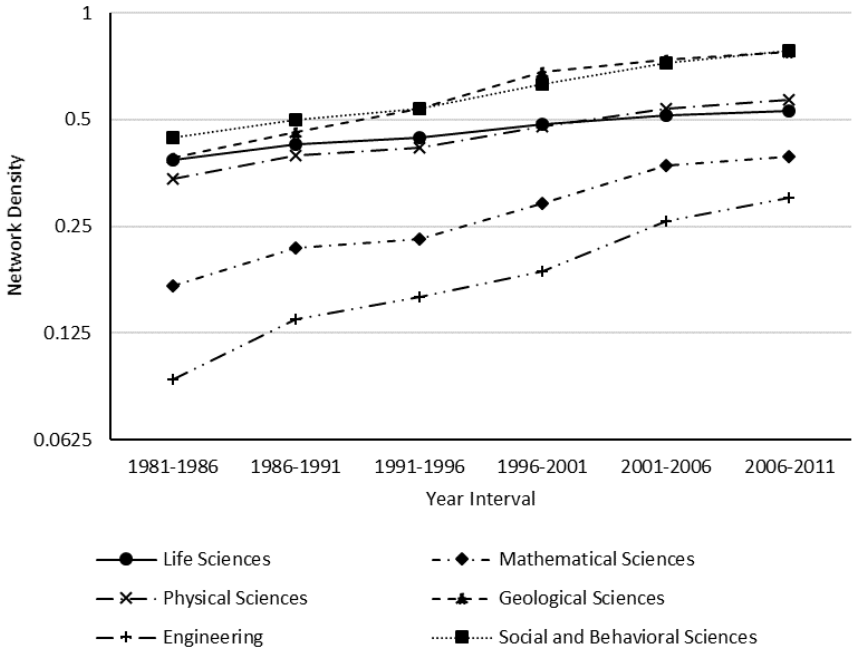


Figure 2
Collaboration Densities Over Time

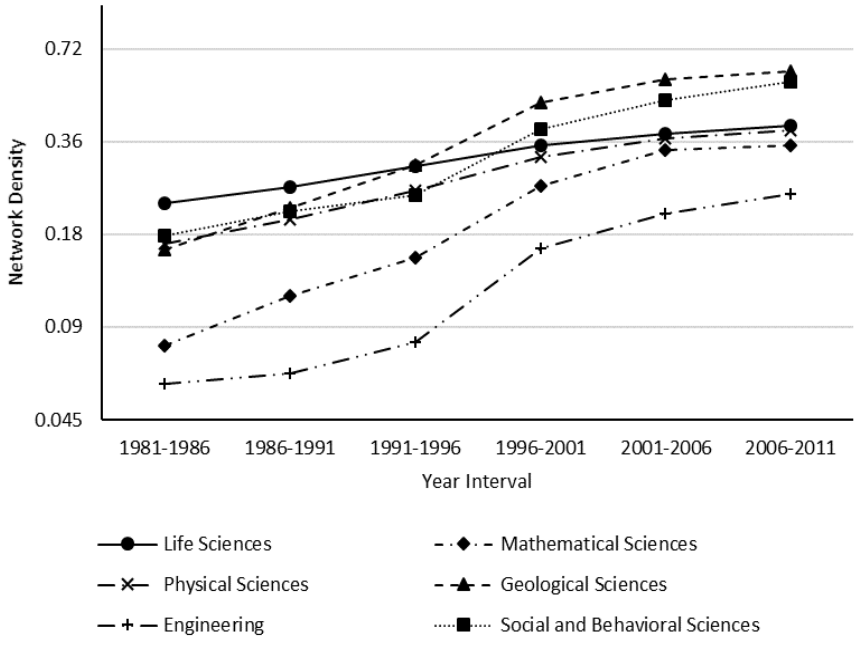


Figure 3A
Federally Funded R&D In Industry and Federal Intramural Defense R&D, 1981-2011, in Millions of 2009 $

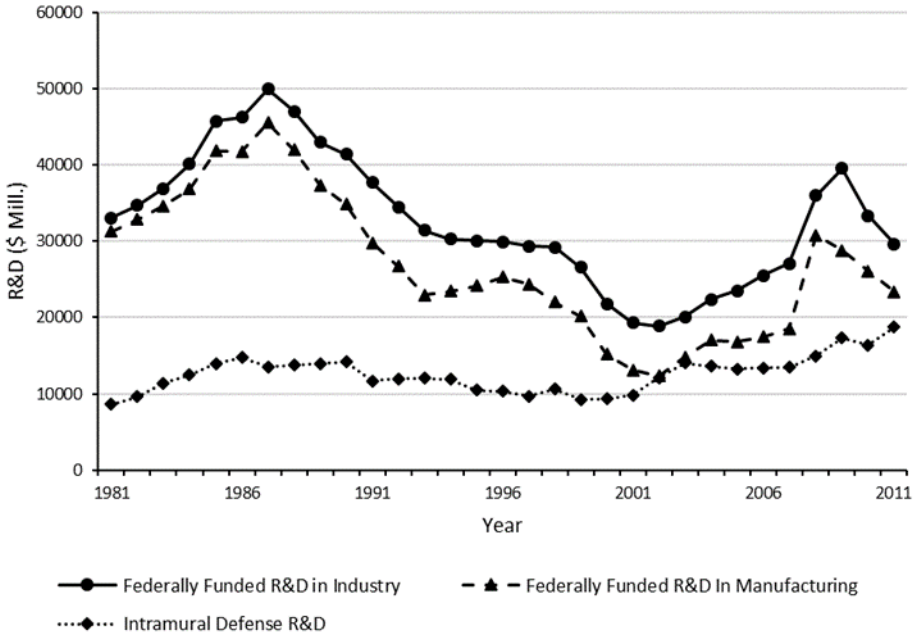


Figure 3B
Federally Funded R&D in Industry and Federal Intramural Defense R&D, 1981-2011 (1981=1.0)

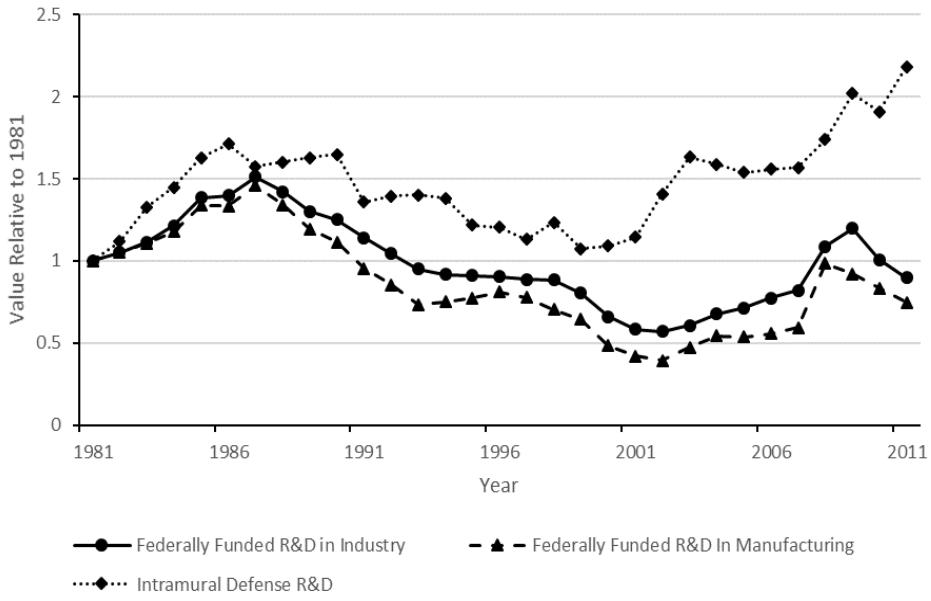


Figure 4A
Federally Funded R&D by Manufacturing Sector,
1981-2007, in Millions of 2009 $

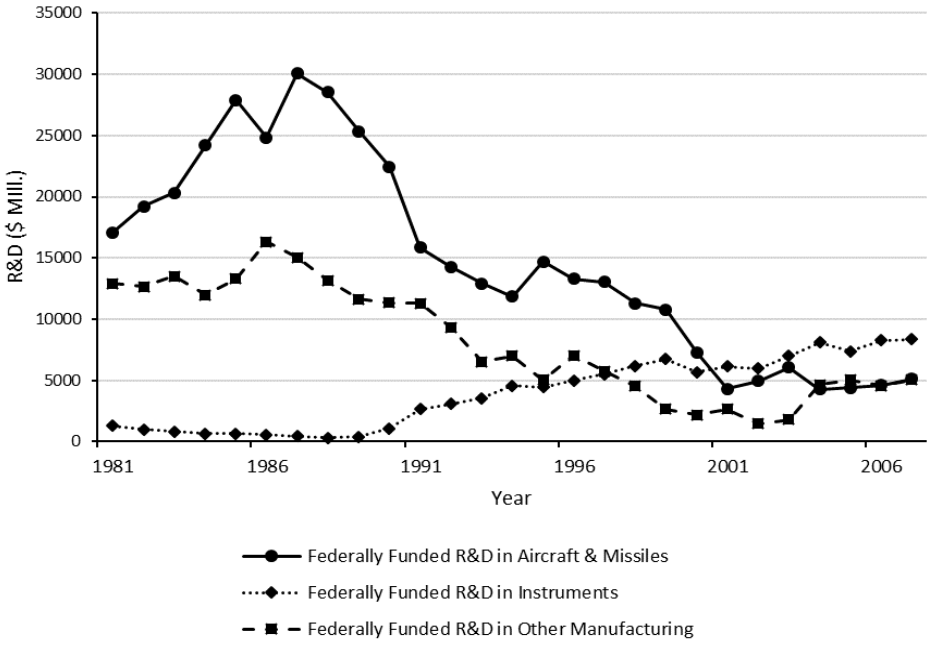


Figure 4B
Federally Funded R&D by Manufacturing Sector,
1981-2007 (1981=1.0)

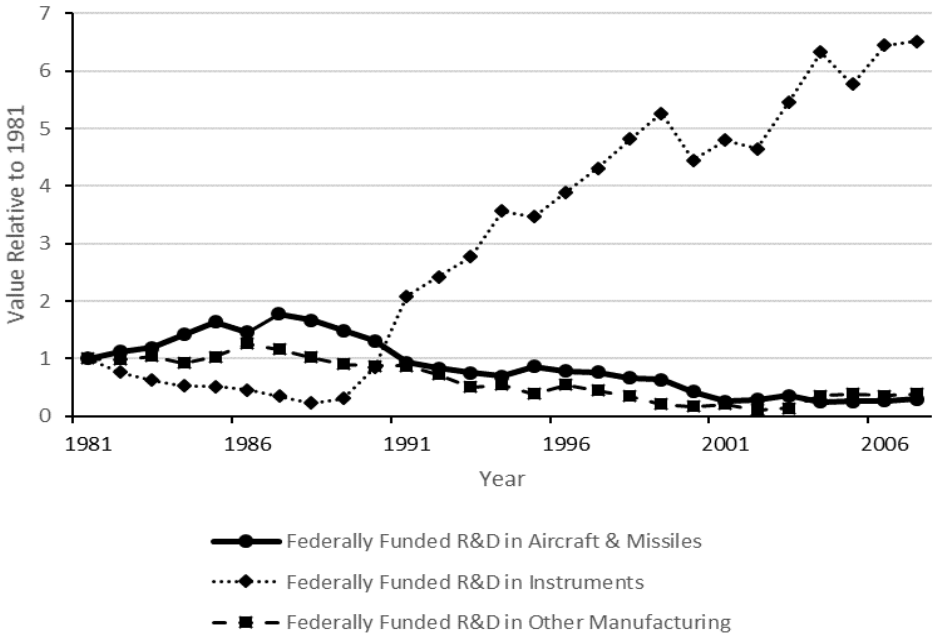


Figure 5A
Federal Support of R&D in Universities, by Groups of Sciences,
1973-2012, in Millions of 2009 $

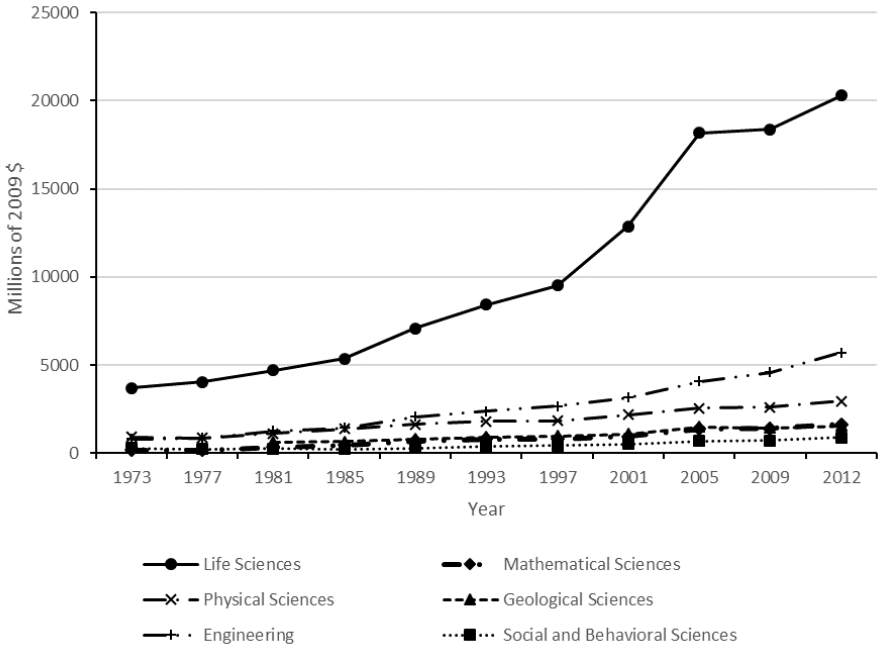


Figure 5B
Federal Support of R&D in Universities, Excluding Life Sciences,
1973-2012, in Millions of 2009 $

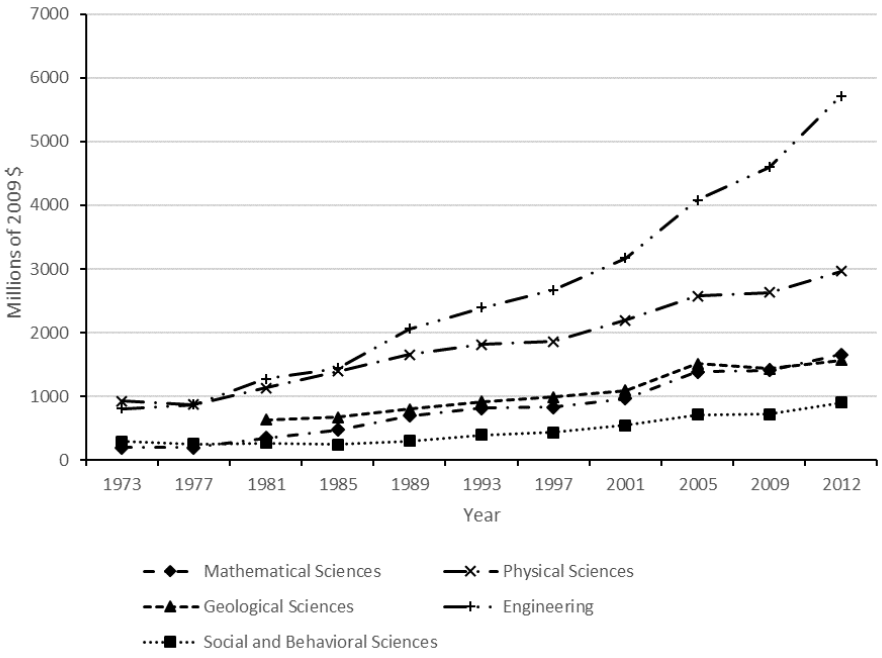
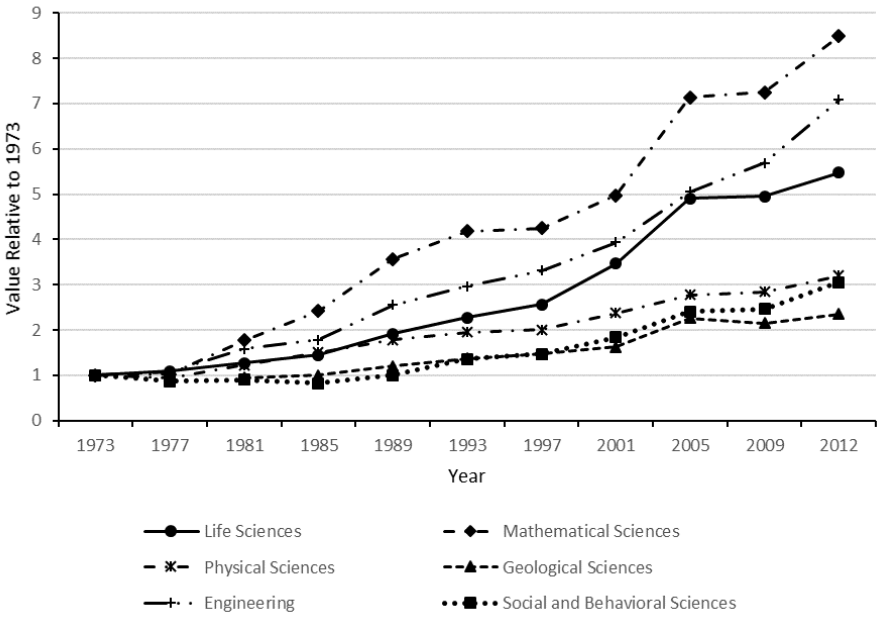


Figure 5C
Federal Support of R&D in Universities, by Groups of Sciences,
1973-2012 (1973=1.0)



On-Line Appendices

Appendix A

Assessing the Scoring Algorithm for Name Disambiguation

In the text we allude to tests of the scoring algorithm for name disambiguation using 5,000 papers written by sixty eminent scientists whose careers are documented in **American Men and Women of Science.** The key advantage of this data is that it reports the career affiliations of scientists of more than average eminence. For each year we know where scientists are employed. We compare publications of ten scientists in the six core fields of biology, chemistry, computer science, engineering, medicine, and physics with produced by the scoring algorithm. To do this, we hand-collected their papers, primarily using *Google Scholar* as the source, and then saved the data in Excel spreadsheets. Matching the research scientists from the two sources was not difficult because we had names and affiliations on which to match. However, a paper may not be present in the hand-collected data but present in the disambiguated data, and vice versa. Only potentially overlapping papers in the two data sources should be considered.

Here are the definitions that we use to compute lumping and splitting error rates for a given scientist, always for the set of papers that overlap. If a paper is correctly assigned by the algorithm, then we increment the *true positive count* by 1. If the paper is incorrectly not assigned to the scientist by the algorithm, then we increment the *false negative count* by 1. Alternatively, if a paper has been incorrectly assigned to the author by the algorithm, then we increment the *false positive count* by 1.

The *lumping error rate* is defined as follows:

(A.1) $$lumping\ error\ rate = \frac{false\ positives}{true\ positives + false\ negatives}$$

The lumping error rate computes the ratio of papers that should not have been assigned to a scientist by the algorithm but were (false positives), to papers that were correctly assigned (true positives) plus papers that should have been assigned to the scientist but were not (false negatives). Consider the papers of a certain Dr. Dumas. The numerator of (A.1) equals papers that the algorithm said were written by Dr. Dumas but were not. The denominator equals papers of Dr. Dumas that were correctly assigned by the algorithm plus papers of theirs that were not assigned by the algorithm but ought to have been. In other

words, the denominator equals the true paper count of Dr. Dumas. The lumping error rate in (A.1) is the number of papers that were incorrectly assigned by the algorithm and were actually someone else's. In other words, (A.1) records lumping mistakes divided by the true paper count for Dr. Dumas.

There is another error rate belonging to Dr. Dumas. The splitting error rate is the number of papers that should have been his, but which the algorithm assigned to someone else, or in other words it is the number of splitting mistakes divided by the true paper count. The *splitting error rate* is:

(A.2) $$splitting\ error\ rate = \frac{false\ negatives}{true\ positives + false\ negatives}$$

Table A1 compiles paper counts and the error rates (A.1) and (A.2) for each of the six fields and across fields. Error rates are weighted: the various paper counts are summed to the field level or across fields and the error rates subsequently computed. The overall lumping error rate is 0.066, or 6.6%, while the splitting error rate is 0.028, or 2.8%. These error rates are quite low given that first names are not available for most scientific papers, and they are not much above the error rates in patent data, where first names are available: see Li (2013).

Table A1
Analysis of Lumping and Splitting Error Rates
Sample of 60 American Men and Women of Science Researchers

Statistic	Engineering	Biology	Chemistry	Medicine	Physics	Computer Science	Total
True Positive Count	326	687	1733	1166	617	408	4937
False Positive Count	0	102	3	161	7	62	335
False Negative Count	12	0	76	1	20	33	142
Lumping Error Rate	0	0.1485	0.0017	0.1380	0.0110	0.1406	0.0660
Splitting Error Rate	0.0355	0	0.0420	0.0010	0.0314	0.0748	0.0280

Source: Authors' calculations, Thomson Reuters Web of Science™, *Google Scholar*, and *American Men and Women of Science* (31st Ed., Gale Cengage, 2013)

Appendix B

Shortest Path Affiliation Algorithm

A. Concepts

Let us say that we are trying to construct the research history of Dr. Dumas. The affiliation algorithm first collects a list of years in which they have been published. For each year, the algorithm collects a list of institutions from their publications. These form candidate institutions for each year. The algorithm then creates a network from these candidate institutions on which Dijkstra's shortest path algorithm will be performed. For example, assume that Dr. Dumas is active from the year 2001 to 2004. In year 2001, they have one publication affiliated to MIT and Harvard and another publication from MIT. In 2002, they have one publication affiliated to MIT and Yale and another to RPI and Yale. In 2003, they have one publication affiliated to Yale and Harvard and in 2004, there is a single publication affiliated to Yale. To summarize, active years and their affiliations are:

- 2001: MIT, Harvard, MIT
- 2002: MIT, Yale, RPI, Yale
- 2003: Yale, Harvard
- 2004: Yale

The candidate graph for Dr. Dumas will be as follows:

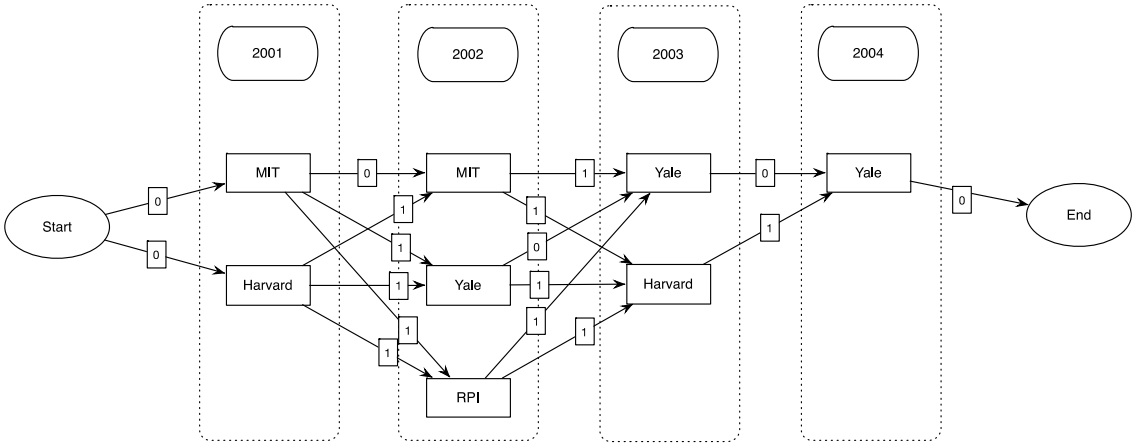


Figure B1: Candidate Graph

This is a network. Every candidate institution in each year is connected by an edge to every candidate institution in the next year. The edge weighs zero if the two institutions are the same, and it weighs 1 if they are different, representing the cost of moving. There is a 'start' and 'end' node in the graph that provides source and destination nodes for the Dijkstra's algorithm. Observe that any path from 'start' to 'end' constructs an affiliation history for Dr. Dumas by selecting one institution every year. It can be easily proved that the shortest path from start node to end will traverse a path in this network that traces year wise affiliation for Dr. Dumas with the least number of moves. For this example, there are three shortest paths, each involving one move:

1. START – MIT – MIT – YALE – YALE – END.
2. START – MIT – YALE – YALE – YALE – END.
3. START – Harvard – YALE – YALE – YALE – END.

The path START – MIT – MIT – YALE – YALE – END corresponds to the affiliation history:

- 2001: MIT
- 2002: MIT
- 2003: YALE
- 2004: YALE

All the shortest length paths are of length one, corresponding to one move, or change in affiliated institution. Dijkstra's shortest path algorithm can choose any one of them. However, we can see that the second path (START – MIT – YALE – YALE – YALE – END) should be preferred since MIT appears twice on the institution list for the year 2001 and Yale appears twice on institution list for the year 2002. We can incorporate frequency of an institution in the affiliation algorithm by reducing edge weights for institutions with higher frequency. More specifically, the weight of incoming edges is decided in the following manner:

- For edge (A, B) if A = B then weight (A, B) = 0.
- If A≠B, then weight (A, B) = 1/free(B)

The candidate graph will now be as follows:

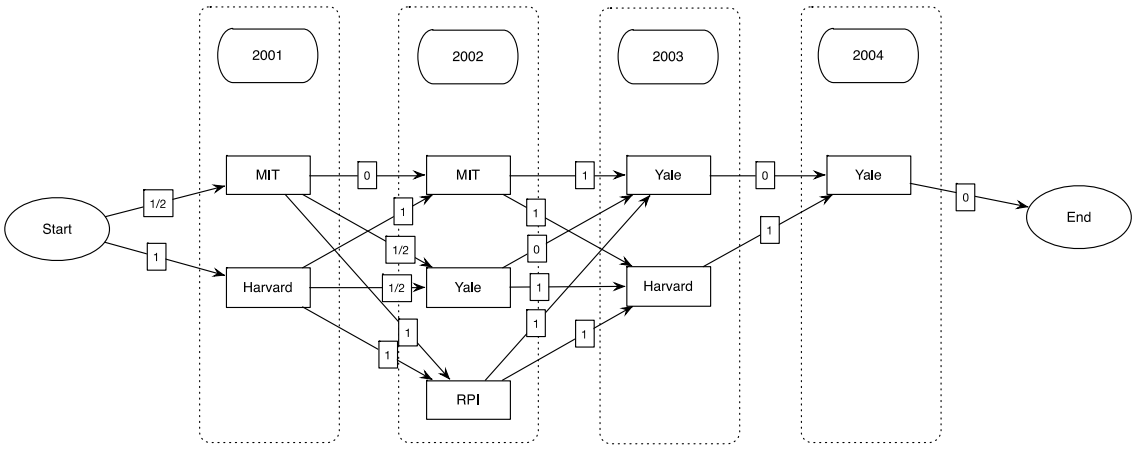


Figure B2: Candidate graph with frequency information

The shortest path algorithm will now choose START – MIT – YALE – YALE – YALE – END since this is the only path with length one. All other paths from START to END have higher weights.

B. Findings using Data from the American Men and Women of Science

Before we analyze results of the affiliation algorithm using the sample of sixty scientists in the **American Men and Women of Science** (AMWS) sample (see Appendix A) we should acknowledge its limitations. The affiliation algorithm uses data generated by the disambiguation algorithm to create a history of institutional affiliations for research scientists. It follows that if there are large lumping or splitting errors for an individual, the affiliation algorithm cannot accurately track their career. The algorithm uses affiliated institutions from published papers as candidate institutions for an individual. If an individual is more prolific, the algorithm has more data points to work with and it can predict the career for the researcher better. However, research scientists with few publications will have low success rates with our algorithm.

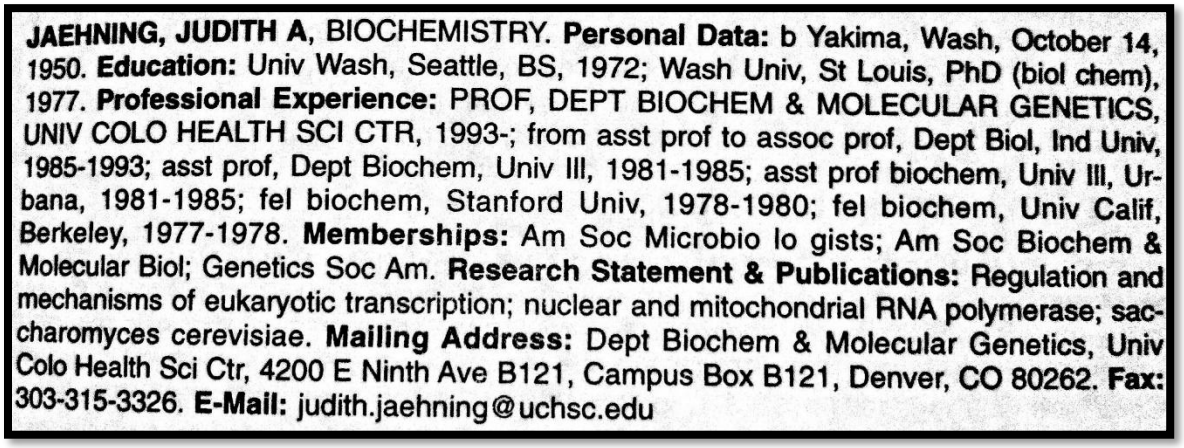
Our research focuses on individuals at top academic institutions, top private firms, and top government research institutions in the U.S. However, the Thompson Reuters data is not limited to this elite group of top institutions. To assign the scope of our study, we first identified papers that had one at

least one top U.S. institutions on their list of affiliations. We then collected all author names from this subset of papers. This became the set of names that were disambiguated. This gives rise to a distinct possibility that authors may not be affiliated with our preferred institutions, but whose coauthors are, will also be part of our disambiguation set. Our affiliation algorithm will not work well on such 'outside' individuals. There will also be individuals who over their careers were associated with top U.S. as well as other institutions. The affiliation algorithm will not be able to track their careers accurately during the time that they were not associated with our top U.S. institutions.

C. Results for Ten AMWS Scientists

This section discusses tests of our affiliation algorithm for ten researchers. It was found that the affiliation algorithm worked well for individuals who never moved or changed institutions, so we focus on individuals that moved one or more times amongst top U.S. organizations and for whom lumping and splitting errors were low. For each of the researchers, we duplicate the AMWS entry if it is available, immediately followed by the sequence of affiliations given by the affiliation algorithm.

1. Jaehning, Judith A



JAEHNING, JUDITH A, BIOCHEMISTRY. **Personal Data:** b Yakima, Wash, October 14, 1950. **Education:** Univ Wash, Seattle, BS, 1972; Wash Univ, St Louis, PhD (biol chem), 1977. **Professional Experience:** PROF, DEPT BIOCHEM & MOLECULAR GENETICS, UNIV COLO HEALTH SCI CTR, 1993-; from asst prof to assoc prof, Dept Biol, Ind Univ, 1985-1993; asst prof, Dept Biochem, Univ Ill, 1981-1985; asst prof biochem, Univ Ill, Urbana, 1981-1985; fel biochem, Stanford Univ, 1978-1980; fel biochem, Univ Calif, Berkeley, 1977-1978. **Memberships:** Am Soc Microbio lo gists; Am Soc Biochem & Molecular Biol; Genetics Soc Am. **Research Statement & Publications:** Regulation and mechanisms of eukaryotic transcription; nuclear and mitochondrial RNA polymerase; saccharomyces cerevisiae. **Mailing Address:** Dept Biochem & Molecular Genetics, Univ Colo Health Sci Ctr, 4200 E Ninth Ave B121, Campus Box B121, Denver, CO 80262. **Fax:** 303-315-3326. **E-Mail:** judith.jaehning@uchsc.edu

name	individual_id	start_year	end_year	pref_name
Jaehning, Judith A	7130364	1984	1988	UNIV ILLINOIS URBANA CHAMPAIGN
Jaehning, Judith A	7130364	1988	1994	INDIANA UNIV
Jaehning, Judith A	7130364	1994	2010	UNIV COLORADO

In this case the affiliation algorithm accurately captures career changes for the researcher from UIUC to Indiana University and then to the University of Colorado. According to the AMWS data (see the upper box), Judith Jaehning was an assistant professor at UIUC from 1981 to 1985, but our algorithm determined the affiliation to be from 1984 to 1988. Such errors could occur for the reasons listed below. It is possible that Jaehning did not start publishing till sometime after 1984 and was still coauthoring papers with other researchers at UIUC after she left for Indiana University. There is also the matter of the publication lag, where a paper is published months or years after submission.

2. Norton, Jack Richard

Norton, Jack Richard
CHEMISTRY.
Personal Data:
b Dallas, Tex, May 5, 1945.
Education:
Harvard Univ, BA, 1967.
Stanford Univ, PhD (chem), 1972.
Professional Experience:
Lee lectr, Univ Chicago, 2001.
PROF CHEM, COLUMBIA UNIV, 1997-.
Guggenheim Found fel, 1989.
from assoc prof to prof chem, Colo State Univ, 1979-1997.
Alfred P Sloan Found fel, 1977-1981.
asst prof chem, Princeton Univ, 1973-1979.
postdoctoral fel Chem, Chem Lab, Univ Cambridge, 1972.
Memberships: Fel AAAS; Japan Soc Prom Sci; Am Chem Soc; Royal Soc Chem.

name	individual_id	start_year	end_year	pref_name
Norton, Jack Richard	7147392	1981	2005	COLORADO STATE UNIV
Norton, Jack Richard	7147392	2005	2011	COLUMBIA UNIV

Here we see that the movement has been captured well, but that year of transition is off by 8 years. We can expect such errors to be present in any affiliation algorithm since their accuracy is sensitive to completeness of the input data.

3. Vitus, Carissima Marie

Vitus, Carissima Marie
SCANNING PROBE MICROSCOPY, CORROSION.

Personal Data:
b Youngstown, Ohio, June 9, 1964.

Education:
Youngstown State Univ, BS, 1986.
Purdue Univ, PhD (analytical chem), 1991.

Professional Experience:
ASST SCIENTIST, BROOKHAVEN NAT LAB, 1992-.
fel, WVa Univ, 1991-1992.
teaching asst gen chem, Purdue Univ, 1986-1990.

Memberships:
Electrochem Soc; Am Chem Soc.

name	individual_id	start_year	end_year	pref_name
Vitus, Carissima Marie	7130501	1990	1993	PURDUE UNIV
Vitus, Carissima Marie	7130501	1993	2001	BROOKHAVEN NATL LAB

The affiliation algorithm correctly tracks career changes of a researcher who goes to a national lab. The algorithm ignores the brief stint at West Virginia University though it is a top U.S. institution. Perhaps no papers were published, or any papers had coauthors from Purdue University.

4. Yao, David D

Yao, David D
QUEUEING NETWORKS, DISCRETE-EVENT SYSTEMS.

Education:
Univ Toronto, MASc, 1981, PhD (indust eng & opers res), 1983.

Honors & Awards:
George E Nicholson Jr Mem Award, Opers Res Soc Am, 1983.
Franz Edel-man Award, Inst Opers Res & Mgt Sci, 1999.
Outstanding Tech Achievement Award, IBM Res, 1999.

Professional Experience:
THOMAS ALVA EDISON PROF, COLUMBIA UNIV, 1992-.
Guggenheim Found fel, 1991-1992.
vis prof, Yale Univ, 1991-1992.
vis scientist, IBM TJ Watson Res Ctr, 1990-2003.
PROF, COLUMBIA UNIV, 1988-.
assoc prof systs eng, Harvard Univ, 1986-1988.
asst prof, Dept Indust Eng & Opers Res, Columbia Univ, 1983-1986.

name	individual_id	start_year	end_year	pref_name
Yao, David D	7130515	1985	2011	COLUMBIA UNIV

This researcher may have published papers during his PhD at University of Toronto but since it is not one of our institutions the algorithm cannot track that affiliation. Brief stints (86-88 at Harvard and 91-92 at Yale) are not tracked. We also note that the researcher was a visiting scientist at IBM while at Columbia University. Our algorithm is not designed to track simultaneous affiliations. In such a situation, the algorithm will select an affiliation that is more sustained, which happens to be Columbia University.

5. Safonov, Michael G

Name	individual_id	start_year	end_year	pref_name
Safonov, Michael G	7130237	1981	1983	HONEYWELL INC
Safonov, Michael G	7130237	1983	2011	UNIV SO CALIF

This is an example in which the affiliation algorithm tracks the movement of a researcher from a private firm that is one of our top U.S. institutions.

6. Mcguinness, Deborah Louise

Mcguinness, Deborah Louise
KNOWLEDGE REPRESENTATION.

Personal Data:
b Drexel Hill, Pa, March 28, 1958.

Education:
Duke Univ, BS, 1980.
Univ Calif, Berkeley, MS, 1981.
Rutgers Univ, PhD (description logics), 1996.

Professional Experience:
Bd mem, Semantic Web Science Foundation, as of 2008.
ENDOWED CONSTELLATION CHAIR, TETHERLESS WORLD RES CONSTELLATION, RENSSELAER POLY-TECH INST, 2007-.
PROF, RENSSELAER POLYTECH INST, 2007-.
actg dir & sr res scientist, Knowledge Systs, Artificial Intelligence Lab, Stanford Univ, beginning 2006.
co-dir & sr res scientist, Knowledge Systs, Artificial Intelligence Lab, Stanford Univ, 2004-2006.
assoc dir & sr res scientist, Knowledge Systs Lab, Stanford Univ, 1998-2004.
CHIEF EXEC OFFICER & PRES, MCGUINESS ASSOCS CONSULT, 1998-.
rotation mgr emerging technol, Personal Online Servs, beginning 1996.
res scientist, Knowledge Representation, Artificial Intel Resdept, AT&T Bell Labs, beginning 1986.
res scientist artificial intel & comput environ res, artificial intel & databases, Home Info Systs, 1984-1986.
res scientist videotex, artificial intel &databases, Home Info Systs, 1980-1984.

Name	individual_id	start_year	end_year	pref_name
Mcguinness, Deborah Louise	7130365	1990	2002	AT&T CORP
Mcguinness, Deborah Louise	7130365	2002	2009	STANFORD UNIV
Mcguinness, Deborah Louise	7130365	2009	2011	RENSSELAER POLYTECH INST

The affiliation algorithm accurately tracks the career of the researcher from AT&T Bell Labs to Stanford and then to RPI. The year of movement is slightly off but the vital information about affiliated organizations is preserved. The year during which Deborah McGuiness left employment at AT&T Corp cannot be established from AMWS data, but it is safe to assume that she left AT&T before becoming a Sr. Research Scientist at Stanford.

7. Meyers, Marian Bennett

Meyers, Marian Bennett
GENETICS, DRUG RESISTANCE.

Personal Data:
b New York, NY, February 22, 1938.
m 1963, c 2.

Education:
Barnard Col, AB, 1959.
Univ Md, MS, 1962.
Cornell Univ, PhD (biochem), 1977.

Professional Experience:
STAFF, DEPT MED, CARDIOVASC INST, MT SINAI SCH MED, as of 1998.
RES ASSOC, SLOAN-KETTERING INST, 1981-.
fel, Sloan Kettering Inst, 1979-1981.
fel genetics, Albert Einstein Col Med, 1977-1979.
Queens Col, 1965-1972.
lectr, org chem, Barnard Col, 1963-1964.
res chemist, Merck & Co, 1962-1963.

name	individual_id	start_year	end_year	pref_name
Meyers, Marian Bennett	7130782	1981	2000	MEM SLOAN KETTERING CANC CTR
Meyers, Marian Bennett	7130782	2000	2003	MT SINAI SCH MED
Meyers, Marian Bennett	7130782	2003	2007	NEW YORK UNIV

The affiliation algorithm correctly identifies the institutions for this researcher, but incorrectly records movement in 2003 from Mt. Sinai School of Medicine to New York University. A further investigation revealed that of Meyers' post-2003 papers, more than a few were not associated with Mt. Sinai, even though her biography says that she was still employed there. This is probably because the institution disambiguation algorithm did not identify all name variations for Mt. Sinai. Consequently, the affiliation algorithm records Meyers as moving to New York University, where some of Meyers' coauthors were affiliated. This highlights a limitation of the affiliation algorithm, which is the quality of data input.

8. Peroutka, Stephen Joseph

Peroutka, Stephen Joseph
NEUROLOGY.

Personal Data:
b Baltimore, Md, February 21, 1954.
m 1979.

Education:
Cornell Univ, AB, 1975.
Johns Hopkins Univ, MD, 1979, PhD (pharmacol), 1980.

Professional Experience:
CHIEF MED OFFICER, ZOGENIX INC, 2007-.
pres & chief exec officer, Synergia Pharma Inc, 2003-2005.
vpres clin res, Deltagen Inc, 2001-2002.
Chief Med Officer, Collabra Pharma, 2000-2001.
franchise develop leader, Johnson & Johnson.
founder, pres & chief exec officer, Spectra Biomed Inc, 1993-1997.
dir, Genentech Inc, 1990-1991.
chief, Neurol Ser, Palo Alto VA Hosp, 1988-1990.
chmn, Am Headache Soc.
asst prof, Neural & Pharmacol, Stanford Univ, beginning 1984-1990.
resident fel, neurol, Johns Hopkins Univ, 1981-1984.
intern med, Stanford Univ, 1980-1981.
franchise develop leader, Johnson & Johnson.

Memberships:
Soc Neuroscience; Am Acad Neurol.

name	individual_id	start_year	end_year	pref_name
Peroutka, Stephen Joseph	7130290	1980	1990	JOHNS HOPKINS UNIV
Peroutka, Stephen Joseph	7130290	1990	1992	STANFORD UNIV
Peroutka, Stephen Joseph	7130290	1992	1998	GENENTECH INC
Peroutka, Stephen Joseph	7130290	1998	2004	UNIV N CAROLINA

This researcher has moved numerous times, often to firms outside our top U.S. institutions. Nevertheless, the affiliation algorithm tracks the first three institutions accurately, through 1991. After this Peroutka moves to startup firms that are outside our set of institutions, and the algorithm tracks incorrectly. The erroneous affiliation with North Carolina could reflect coauthor affiliations.

9. Wysocki, Lawrence J

Wysocki, Lawrence J
PEDIATRICS.

Education:
Univ Calif, Irvine, BS, 1975.
Harvard Univ, PhD (biol), 1981.

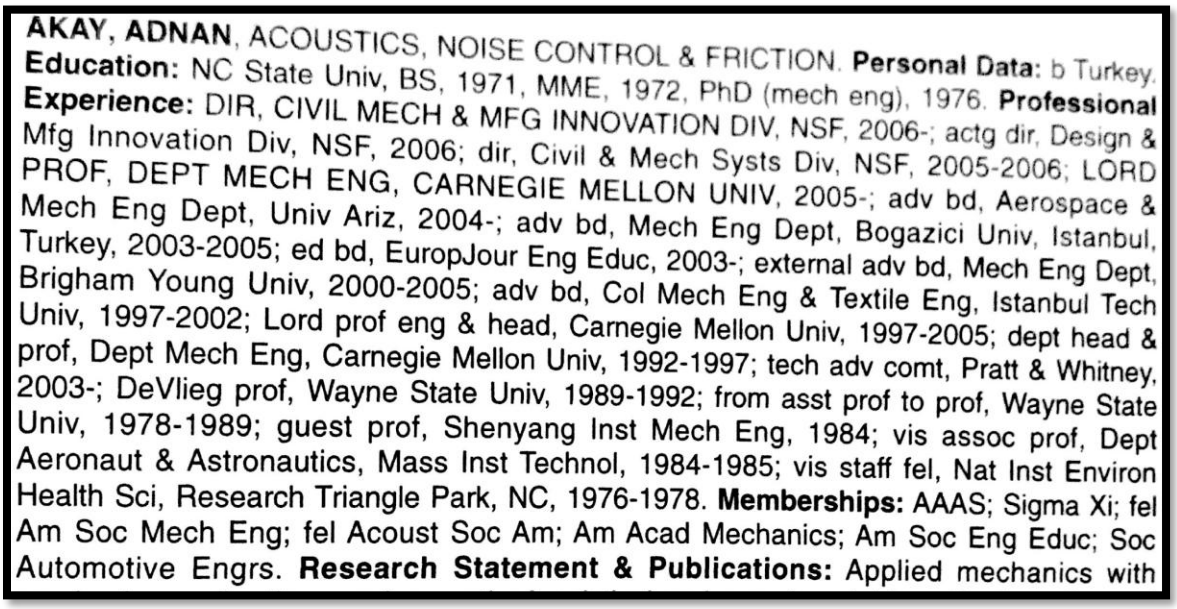
Professional Experience:
PROF, DEPT IMMUNOL, UNIV COLO, DENVER, as of 2008.
PROF, INTEGRATED DEPT IMMUOL, NAT JEWISH MED & RES CTR, 2000-.
assoc fac mem, dept pediatrics, Nat Jewish Med & Res Ctr, 1994-2000.

Research Statement & Publications:
Development of B cell memory in immunity and autoimmunity.

Name	individual_id	start_year	end_year	pref_name
Wysocki, Lawrence J	7130239	1981	1994	HARVARD UNIV
Wysocki, Lawrence J	7130239	1994	2011	UNIV COLORADO

The affiliation algorithm correctly identified the movement of this researcher from Harvard to the University of Colorado. Since the National Jewish Med & Res Ctr was not part of our preferred list of institutions, it was not tracked.

10. Akay, Adnan



AKAY, ADNAN, ACOUSTICS, NOISE CONTROL & FRICTION. **Personal Data:** b Turkey. **Education:** NC State Univ, BS, 1971, MME, 1972, PhD (mech eng), 1976. **Professional Experience:** DIR, CIVIL MECH & MFG INNOVATION DIV, NSF, 2006-; actg dir, Design & Mfg Innovation Div, NSF, 2006; dir, Civil & Mech Systs Div, NSF, 2005-2006; LORD PROF, DEPT MECH ENG, CARNEGIE MELLON UNIV, 2005-; adv bd, Aerospace & Mech Eng Dept, Univ Ariz, 2004-; adv bd, Mech Eng Dept, Bogazici Univ, Istanbul, Turkey, 2003-2005; ed bd, EuropJour Eng Educ, 2003-; external adv bd, Mech Eng Dept, Brigham Young Univ, 2000-2005; adv bd, Col Mech Eng & Textile Eng, Istanbul Tech Univ, 1997-2002; Lord prof eng & head, Carnegie Mellon Univ, 1997-2005; dept head & prof, Dept Mech Eng, Carnegie Mellon Univ, 1992-1997; tech adv comt, Pratt & Whitney, 2003-; DeVlieg prof, Wayne State Univ, 1989-1992; from asst prof to prof, Wayne State Univ, 1978-1989; guest prof, Shenyang Inst Mech Eng, 1984; vis assoc prof, Dept Aeronaut & Astronautics, Mass Inst Technol, 1984-1985; vis staff fel, Nat Inst Environ Health Sci, Research Triangle Park, NC, 1976-1978. **Memberships:** AAAS; Sigma Xi; fel Am Soc Mech Eng; fel Acoust Soc Am; Am Acad Mechanics; Am Soc Eng Educ; Soc Automotive Engrs. **Research Statement & Publications:** Applied mechanics with

name	individual_id	start_year	end_year	pref_name
Akay, Adnan	7130350	1981	1996	WAYNE STATE UNIV
Akay, Adnan	7130350	1996	2012	CARNEGIE MELLON UNIV

The affiliation algorithm accurately tracks this researcher's movement from Wayne State University to Carnegie Mellon University. According to the AMWS biography, Akay served on the 'Adv Bd' of many colleges during his career. It is not clear what this position entails but it could be assumed that he did not perform research at those institutions.

Appendix C

Interior Solution of the Problem of the Individual Scientist

The first order conditions for a maximum of (1) in Section II.A are:

(A.3) $$\frac{\partial Y_i}{\partial E_{Ai}} = \mu \cdot W_A \cdot A_i \cdot E_{Ai}^{\mu-1} \cdot \lambda_i^{\sigma} - \mu \cdot W_R \cdot R_i \cdot (E_i - E_{Ai})^{\mu-1} \cdot \lambda_i^{\sigma} = 0$$

(A.4) $$\frac{\partial Y_i}{\partial \lambda_i} = \sigma \cdot W_A \cdot A_i \cdot E_{Ai}^{\mu} \cdot \lambda_i^{\sigma-1} - \sigma \cdot W_R \cdot R_i \cdot (E_i - E_{Ai})^{\mu} \cdot \lambda_i^{\sigma-1} = 0$$

Simplifying (A.3),

(A.5) $$\frac{E_{Ai}}{E_i - E_{Ai}} = (a_i)^{\frac{1}{1-\mu}} \cdot \left(\frac{\lambda_i}{1-\lambda_i}\right)^{\frac{\sigma}{1-\mu}}$$

Recall that $a_i \equiv \frac{W_A \cdot A_i}{W_R \cdot R_i}$ equals the value weighted relative stock of administrative skill. Likewise, simplifying (A.4) yields

(A.6) $$\frac{\lambda_i}{1-\lambda_i} = (a_i)^{\frac{1}{1-\sigma}} \cdot \left(\frac{E_{Ai}}{E_i - E_{Ai}}\right)^{\frac{\mu}{1-\sigma}}$$

Substitute (A.6) into (A.5), thereby eliminating $\frac{\lambda_i}{1-\lambda_i}$, and solving for administrative effort E_{Ai}. The result is:

(A.7) $$E_{Ai} = \frac{(a_i)^{\frac{1}{1-\mu-\sigma}} \cdot E_i}{1 + (a_i)^{\frac{1}{1-\mu-\sigma}}}$$

In (A.7), $\frac{1}{1-\mu-\sigma} > 0$ since $0 < \mu + \sigma < 1$ by diminishing returns. It follows that $0 < E_{Ai} < E_i$. Using the effort constraint, $E_{Ri} = E_i - E_{Ai}$, we reach the solution for research effort:

(A.8) $$E_{Ri} = \frac{E_i}{1 + (a_i)^{\frac{1}{1-\mu-\sigma}}}$$

Substitution of (A.7) and (A.8) into (A.6) reveals that:

(A.9) $$\lambda_i = \frac{(a_i)^{\frac{1}{1-\mu-\sigma}} \cdot E_i}{1 + (a_i)^{\frac{1}{1-\mu-\sigma}}}$$

(A.10) $$1 - \lambda_i = \frac{E_i}{1 + (a_i)^{\frac{1}{1-\mu-\sigma}}}$$

and proves equations (11) and (12) of Section V.A.

Appendix D

Table D1
Best Response Demand for Research Scientists
Citation and Collaboration In-Degree Centralities
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(D1.1)	(D1.2)	(D1.3)	(D1.4)	(D1.5)	(D1.6)
Dependent Variable	All Research Scientists			Long-Term Research Scientists		
Time Period Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Science Field Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Institution Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Method	OLS	2SLS	2SLS	OLS	2SLS	2SLS
VCE Assumption	Robust	Robust	Clustered [a]	Robust	Robust	Clustered [a]
Citation In-Degree Centrality	0.114**	0.192**	0.192**	0.084**	0.142**	0.142*
	(0.004)	(0.035)	(0.081)	(0.004)	(0.029)	(0.068)
Collaboration In-Degree	0.533**	-1.168	-1.168	0.012	-1.418**	-1.418
Centrality	(0.073)	(0.657)	(1.521)	(0.065)	(0.538)	(1.247)
Number of Observations	50,615	50,615	50,615	50,615	50,615	50,615
F (286, 50,328)	527.16			343.48		
Probability>F	0.0000			0.0000		
Wald Chi2(286)		112,196.6	33,573.2		75,387.3	115,538.0
Probability>Chi2		0.0000	0.0000		0.0000	0.0000
R-squared	0.930	0.901	0.901	0.844	0.799	0.799
Root MSE	257.96	304.88	304.88	260.11	294.16	294.16

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data cover the years from 1986-2011. The period 1981-1986 is deleted, since IVs are lagged one period. See the text and Table 3 for a description of the covariates. This table omits all covariates except the centralities. Equations are specified as in equations 5.2-5.3 or 5.5-5.6 of Table 5 in the text. [a] Clustering is by institution and field. **Coefficient differs from zero at the one percent level. *Coefficient differs from zero at the five percent level.

Appendix E

Table E1
Best Response Demand for All Research Scientists
By Field, Citation In-Degree Centralities
2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Field of Science	Centrality Coefficient (Clustered Standard Error [a])
Aeronautical Engineering	-1.122
	(1.59)
Agriculture	0.072
	(0.055)
Astronomy	0.037*
	(0.020)
Atmospheric Sciences	0.068***
	(0.016)
Biology	0.165***
	(0.032)
Chemical Engineering	0.320
	(0.255)
Chemistry	0.076***
	(0.013)
Civil Engineering	0.389
	(0.286)
Computer Science	0.438*
	(0.256)
Geology	0.184**
	(0.079)
Economics	-0.027
	(0.099)
Electrical Engineering	-0.509
	(0.975)
Materials Science Engineering	0.060
	(0.136)
Mathematics and Statistics	0.212***
	(0.079)
Mechanical Engineering	1.456
	(1.299)
Medicine	0.093
	(0.070)
Oceanography	0.096
	(0.061)
Other Engineering	0.386***
	(0.106)

Table E1
Best Response Demand for All Research Scientists
By Field, Citation In-Degree Centralities
2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Field of Science	Centrality Coefficient (Clustered Standard Error [a])
Physics	0.021
	(0.057)
Psychology	0.002
	(0.059)
Binomial Test that Coefficient Sign is Random ($\theta_0 = 0.5$) [b]	
Probability that at least 17 of 20 coefficients are positive	0.00129***
Probability that more than 17 of 20 coefficients are positive	0.00020***

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. The period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates. This table omits all covariates except the centralities. Equations are specified as in Table 5, Equation 5.3 of the text. [a] Clustering is by institution and field. [b] θ is the probability of success in the binomial distribution, where success in this case means a positive sign of the centrality coefficient. The null hypothesis that $\theta_0 = 0.5$ is tantamount to an assumption that the coefficient sign is random. *** Coefficient differs from zero at the one percent level. ** Coefficient differs from zero at the five percent level. * Coefficient differs from zero at the ten percent level.

Table E2
Best Response Demand for All Research Scientists
By Groups of Fields, Citation In-Degree Centralities
2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Field of Science	Centrality Coefficient (Clustered Standard Error [a])
Engineering	0.289
	(0.205)
Life Sciences	0.127***
	(0.020)
Mathematical Sciences	-0.253
	(0.558)
Physical Sciences	0.128***
	(0.050)
Earth Sciences	0.128**
	(0.055)
Social and Behavioral Sciences	0.096***
	(0.022)
Binomial Test that Coefficient Sign is Random ($\theta_0 = 0.5$) [b]	
Probability that at least 5 of 6 coefficients are positive	0.10937*
Probability that all 6 coefficients are positive	0.01562**

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. The period 1981-1986 is deleted since IVs are lagged one period. See the text and Table 3 for a description of the covariates. This table omits all covariates except the centralities. Equations are specified as in Table 5, Equation 5.3 of the text. [a] Clustering is by institution and field. [b] θ is the probability of success in the binomial distribution, where success in this case means a positive sign of the centrality coefficient. The null hypothesis that $\theta_0 = 0.5$ is tantamount to an assumption that the coefficient sign is random. Engineering includes aeronautical engineering, chemical engineering, civil engineering, electrical engineering, materials science engineering, mechanical engineering, and other engineering. Life sciences include agriculture, biology, and medicine. Mathematical sciences include mathematics and statistics and computer science. Physical sciences include astronomy, chemistry, and physics. Earth sciences include atmospheric sciences, geology, and oceanography. Social and behavioral sciences include economics and psychology. *** Coefficient differs from zero at the one percent level. ** Coefficient differs from zero at the five percent level. * Coefficient differs from zero at the ten percent level.