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

March 2026: Revision 1

Comments Welcome

Keywords: Social and Economic Networks, Linear Model of Social Interactions, Demand for Scientists and Engineers, Network Centralities, Research & Development, Defense Procurement, Government Policy, Strategic Complementarity

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.

** We thank the National Bureau of Economic Research for its support of our research. Clarivate Analytics advised us concerning the Web of Science™, which forms a large part of our data. Malik Magdon-Ismail, Professor of Computer Science at Rensselaer Polytechnic Institute, provided advice as regards the algorithms essential for identifying the scientists and engineers in this paper. We thank Eli Berman, Ian Cockburn, Robin Cowan, Julie Hotchkiss, Adam Jaffe, Megan Macgarvie, Peter Orazem, Scott Stern, and participants at the Midwest Economic Association Meetings, the annual Social Science Conference at the U.S. Military Academy at West Point, the NBER Productivity Seminar, and the 2025 Workshop on the Organization, Economics, and Policy of Scientific Research, for valuable criticisms. NSF award SMA-1064227 helped fund this project.

Abstract

Using Web of Science™, we construct citation networks for U.S. research institutions during 1981-2011. Machine learning identifies nodes consisting of institutions, locations and fields. Linked by citation edges, the nodes form networks of scientists and scientific research.

Edges define centralities that in turn determine importance and size of nodes. In particular, In-Degree and PageRank centralities help govern the demand for scientists and the supply of scientific research. With randomly assigned centralities as instruments, we find that, particularly in universities, centralities increase scientific employment, indicating that research in one node is a strategic complement to research elsewhere. At the same time, scientific research captured by future citations rises with the square of scientific employment, indicating large payoffs from research in influential nodes. Thus, government policy creates a system of positive externalities from scientific research.

In the next draft of this paper, Friedman regressions, named after Milton Friedman, will study the relationship between federal funding of university research and the citation centralities. If funding is efficiently distributed, then, holding constant field of science and thus intrinsic differences in field contributions, funding should rise in proportion with importance and hence the centralities. But if funding is inefficient as Friedman asserted, then funding should rise less than proportionately with the centralities.

I. Introduction

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

This paper finds that social networks of scientists and engineers sustain these outcomes. Government support has created a network of 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]. The nodes are likewise 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 the 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 government 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 V is that scientific research in one node increases research in other

[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.

nodes and does form a strategic complement to that research. Therefore, government 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. The networks contain 255 top U.S. research institutions that account for nearly all scientific research in the country[4]. Up to twenty fields of science occur in an institution-location[5]. 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[6].

Section II presents an economic analysis of networks. It models the behavior of nodes of the network, based on a version of the linear model of social interactions. 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, and decreases it if they are strategic substitutes. In this model the maximum value function of research production depends on squared employment of research scientists, contingent on best response being satisfied. This result indicates large payoffs from research by larger and more influential nodes.

Section III initiates the empirical work whose goal is to estimate the model of Section II. Section III is specifically concerned with data identification, a step that is required by the available data. Because this

[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 network data.

[5] The twenty fields follow the system of classifying fields of the U.S. National Science Foundation. This system is more detailed for engineering disciplines than for example, the life sciences, which consist of agriculture, biology, and medicine. Given the large scale of these fields, their division into subfields is a useful frontier for future development by NSF statisticians. In our explorations of the institutional data, we were unable to devise specific words that would classify, for instance, medical graduate programs into consistent subfields. These same points apply to mathematics and statistics, chemistry, physics, and other fields. The detail assigned to non-engineering fields is not as fine as for engineering subfields.

[6] 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.

step is unfamiliar, we provide an overview here, with much more detail contained in a related paper (Adams, Haté, and Clemmons, 2026). Section III.A identifies U.S. institutions using a list of alternative names for each institution. Clarivate Analytics provided us with unifying name files for nonprofit organizations. However, these files are not available for firms. In this case we identify the top 100 R&D firms 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 with Top 100 R&D firm names in Compustat and subsidiary names of the firms in Mergent Online. The result is a set of items with aliases of Top 100 R&D firms defined as Web of Science institutions.

Section III.B identifies individual scientists, distinguishing them from author names. This step relies on iterative scoring methods that weigh author and article information to determine whether an author's name represents the same person across scientific papers. The work assigns 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. Computationally a bucket is a community of non-overlapping identities who share the same name-hash but are different scientists. To split the community into distinct identities we apply a widely used algorithm. This is the Fast Modularity Community Structure Inference Algorithm due to Clauset, Newman, and Moore (2004).

Before applying the algorithm, we cluster author names on different papers using 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 on the similarity score is needed to split the graph for each bucket into separate groups. The Fast Modularity algorithm sets this lower cutoff.

We manually evaluate the scoring mechanism twice. First, we evaluate the scoring mechanism on a small sample to identify any shortcomings. We then revise the initial scoring mechanism. Second, we use a sample of sixty eminent scientists from **American Men and Women of Science** whose papers were known to us in advance. We compare predicted to known authorship to assess the revised scoring

mechanism. We find errors of a few percentage points that are comparable to prior research on patents. Details are in Section III.B. Only then do we apply the scoring mechanism and Fast Modularity Community algorithm to the entire dataset.

Section III.C assigns scientists from 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 is this. Since moving between jobs is costly, we use Dijkstra's Algorithm to minimize the number of employment changes. We evaluate the algorithm by comparing actual to predicted changes using the sample of sixty eminent scientists. Most scientists rarely change affiliations and errors in changes in employment changes are small. A more important point is that errors in the timing of employment change are independent. These errors cancel out when we aggregate scientists by node and year, as we do in the empirical work.

Given these steps, we find that there are 5,850,091 scientists in U.S. institutions. This is 3.4 scientists per bucket[7]. There are 6,350,323 scientist-affiliations, or 1.1 affiliations per scientist. This concludes the data identification steps in Section III.

Section IV constructs counts of researchers, papers, and citations received by node and year, where nodes are institutions, locations, and fields[8]. Combined with citations between nodes, this section defines contents of the nodes forming the citation networks.

For tractability we break the networks up into six-year intervals from 1981-1986, 1986-1991, …, 2006-2011. To avoid small nodes for which scientific research is inherited from prior graduate work, we impose cutoffs. First, we require that university-field nodes have graduate programs and that they publish at least sixty papers during each six-year interval[9]. Second, for non-university nodes during each six-year

[7] Since there are 5,850, 091 scientists and 1,732,238 buckets, this is 3.4 scientists per bucket.
[8] Fractional papers written at an institution, location, field and year are the share of its authors on each paper summed over all papers for the node and year. This is the same for future citations received. The use of fractional papers rather than whole papers avoids double counting of papers and citations.
[9] We experimented with a lower cutoff of thirty papers on universities with comparable results. Including these smaller departments made slight difference to the estimates.

interval, we require at least four hundred publications in an institution and field, and at least one hundred papers in an institution-field-location.

We calculate two centrality measures using the citation networks. These are citation in-degree centrality and citation PageRank centrality[10]. There are advantages and disadvantages to each measure. The in-degree centrality retains significant differences between fields and institutions in the number of incoming citations. And of course, network scientists prefer in-degree centrality to out-degree centrality because it indicates influence over neighboring nodes.

The PageRank centrality also has advantages. It is readily computed on citation networks. 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.

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 the effect of normalization is to flattens field and institution effects compared to the in-degree centrality.

Building on the model in Section II and the subsequent work of data construction, Section V 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

[10] We compute other measures, but we focus on in-degree and PageRank centralities because of their attractive properties.

the degree centrality and omitted variables that could 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[11]. They are 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

[11] 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.

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[12]. For references to recent literature, see Jackson, Rogers, and Zenou (2017)[13].

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 presents the linear-quadratic model of nodes in scientific networks. Section III conducts data identification steps needed for citation networks composed of scientific institutions, locations, and field nodes. Section IV builds networks of citations at the node level and divides the data into separate networks lasting six years over 1981-2011. Section V estimates demand functions for research scientists and supply equations for research output. Section VI summarizes, concludes, and suggests further directions for research.

[12] 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.

[13] 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. Behavior of Nodes of Science and Engineering Networks

A. 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 (or teaching) and research over administrative and research scientists in the labor force. We assume 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[14]. 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 included in the model. 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 $p_A q_{Ai}$ as a linear-quadratic function of administrative scientists ℓ_{Ai}, and likewise for the value of research

[14] 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.

output $p_R q_{Ri}$ and research scientists ℓ_{Ri} . Prices and quantities of administration and research are p_A and p_R and q_{Ai} and q_{Ri} respectively. The node's problem is:

$$(1) \qquad \underset{\ell_{Ai},\ell_{Ri}}{max} \begin{Bmatrix} NV_{Ai} \equiv p_A q_{Ai} - \frac{\eta_A}{2}\ell_{Ai}^2 = v'_{Ai}\varphi_A + \beta_{Ai}\ell_{Ai} - \frac{\eta_A}{2}\ell_{Ai}^2 + e_{Ai} \\ NV_{Ri} \equiv p_R q_{Ri} - \frac{\eta_R}{2}\ell_{Ri}^2 = v'_{Ri}\varphi_R + \beta_{Ri}\ell_{Ri} + \gamma_C\left(\sum_j C_{ji}\ell_{Rj}\right)\ell_{Ri} \\ +\gamma_T\left(\sum_j T_{ji}\ell_{Rj}\right)\ell_{Ri} - \frac{\eta_R}{2}\ell_{Ri}^2 + e_{Ri} \end{Bmatrix}$$

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

$$(2) \qquad \beta_{Ai} = x'_{Ai}\theta_A + u_{Ai}$$

$$(3) \qquad \beta_{Ri} = x'_{Ri}\theta_R + u_{Ri}$$

From equation (1), $p_A q_{Ai} = v'_{Ai}\varphi_A$ and $p_R q_{Ri} = v'_{Ri}\varphi_R$ 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[15]. Therefore, following Manski (1993), network effects are exogenous or contextual effects that interact with network centralities.

In the bottom equation of (1), 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 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 (1), for administration or teaching, lacks network terms. It conforms to our view that knowledge flows and resource sharing are more important for research than for administration. However, networks of contacts could be keys to administration. Alternatively, research networks could

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

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:

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

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:

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

Equation (5) 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 (5) into the research value equation at the bottom of (1) yields

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

Since p_R is unknown we normalize (6) by this output price. The empirical work concentrates on the best response demand equation (5) for research scientists and the optimal supply of research output (6).

To summarize, implications of this section are as follows. First, network effects tend to be concentrated on research scientists. Second, in theory, there are two network effects, one from citations, the other from collaborations. But in practice, a separate effect of collaboration is hard to distinguish owing to collinearity of the centralities. Third, if strategic complementarity prevails, network effects increase research scientists on the extensive margin of an institution, location, and field. Fourth, if there is a behavioral response from individual scientists to greater rewards for research, this leads to a reallocation towards research on the intensive margin of the individual scientist. In that case network effects could enter negatively in the equation for administrators or teachers if administrative labor were observable.

III. Data Identification

This section undertakes three kinds of data identification. It determines institutions, scientists, and institutional affiliations. For this purpose, we apply machine learning to the Web of Science™ database[16]. This is possible because each paper is associated with author names, addresses, institutional affiliations, journal names, journal classifications by science field, and publication dates that aid in data identification.

As noted, there are three sources of ambiguity. First, organization names conceal 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 IV undertakes these steps.

A. Assignment of Research Organizations to Institutions

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

Organization names are variants that are abbreviations and alternate names. Fortunately, Thomson-Reuters provided unification files for U.S. nonprofit institutions, that is, all of our institutions beside firms, which contain name variants, allowing us to assign items to parent U.S. nonprofits.

[16] 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. For reasons of familiarity in this section, we address the company using the original parent, Thomson-Reuters even though the firm and division have since parted ways.

Thomson-Reuters provided 407 unification files that represent more institutions than we ultimately include in the study, because some institutions are small and publish too intermittently to be retained for analysis. The files contain six columns. 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. An additional seventh column contains entries that are used if component names are ambiguous. Using these data, we assign organization names to U.S. nonprofit research institutions.

To accomplish this same goal for parent U.S. firms, we extract names of the Top 100 R&D firms in the 2012 edition of Compustat, at the outset of the project, and we collect names of subsidiaries of these parent firms that are listed under Mergent Online™. Using subsidiary names, we construct organizational directories for firms that we match for institutions in The Web of Science™.

B. Disambiguation of Author Names

The next step determines which author's names, given similar names on different papers, stand for the same person. This step establishes the identity of scientists in this study. This process is known as entity resolution or disambiguation in computer science[17]. Authors sharing the same last name and first initial, known as a name hash, are treated as potentially the same author, 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. The largest bucket contains an astounding 24,332 entries. To reveal identities of persons 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 mechanisms, our algorithm is developed using training data sets for which identities are known in advance. Total scores are defined for pairs of publications based on the following seven scores.

[17] 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 find that our machine learning approach is accurate within a few percentage points.

1. Author Score

Author score uses first name, second initial, suffix, and author key words for matching names across publications. If first names are not present in a pair of 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 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. Thus, 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 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 must account for frequently occurring word overlaps that do not imply similarity. We discard words such as ‘sci,’ ‘USA,’ ‘japan,’ ‘hosp,’ ‘england,’ and so on. We also discard articles and stop words, which 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

again 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 scores adjust this 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, where each group represents 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 tests, we find that a cutoff on the similarity score of at least 20 out of 80 points fits the data best and identifies any two publications as assigned to the same person.

We apply our algorithm to 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)[18]. 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 scientific papers mostly do not, our algorithm can be said to perform roughly comparable 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. We assign authors to the most common field listed in their publications. But each paper is assigned to the field that represents the majority of authors, and hence the audience for the paper. This is a major step forward since it replaces crude assignments of publications according to most common field for a journal with a more nuanced assignment that reflects

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

interests of individual authors. To see the advantage of this, consider interdisciplinary journals. Under our scheme, an article in **Science** or **Nature** written by a group of physicists is classified 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[19].

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 of institutional affiliations over time. This section provides an overview of the affiliation algorithm that makes this feasible.

Changing institutions requires 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 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[20]. 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

[19] 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.
[20] The data are too voluminous to identify institutional affiliations through manual checking.

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 cancels out errors of dating. All the steps discussed in this section are addressed in a companion descriptive paper by Adams, Haté, and Clemmons (2026).

IV. Constructing the Networks

A. Nodes of the Networks

Using the results of Section III we collect scientists, fractional papers, and fractional citations received over the next five years by institutions, locations, fields, and year[21]. Together these data cover the evolution of citation networks during 1981-2011. Since prior steps have associated individual scientists with an institution and field at different points in time, henceforth our discussion is concerned with nodes and edges of the networks.

Summing scientists by node and year yields the number of research scientists. Since we know the prior years that a scientist publishes with a node, we can associate duration of publication at a node with 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 durations of six years or more, which we call long-term research scientists. By node and year, we also accumulate fractional papers and fractional citations over the next five years.

Permanence of a node is desirable, because it indicates ongoing research. And yet many nodes are small, publish occasionally, and only intermittently belong to networks. These are unlikely to form stable research groups. Across the sciences we find that 1,793 U.S. university nodes, or departments, have

[21] 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.

graduate programs and receive federal funding. Since graduate teaching and the pursuit of funding involve fixed costs these are likely permanent research programs. The 1,793 departments capture all major research groups in U.S. universities.

Besides universities we identify 892 institution-location nodes for firms, government agencies, government laboratories, and biomedical research centers in 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 therefore 2,685 (1,793+892), of which about two-thirds are universities and one-third are other institutions. Not all nodes exist at every point in time so 2,685 is an upper bound on the number of nodes in any year. These restrictions correspond to 255 institutions. This set of institutions dominates U.S. scientific research.

B. Edges of the Networks

Consider the citation edges that connect the nodes. We are primarily interested in edges of citation networks; 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. Appendix A contains more information on this point.

Nevertheless, for clarification we distinguish properties of the two networks. Citations are more abundant than collaborations. Citations result from searching the literature searches, subject to constant marginal costs. Collaborations are costly; their marginal cost rises with distance between collaborators, and with the number of collaborations. Edges follow different rules in the two networks. Citations are controlled by citing scientists. Collaborations require assent of scientists in the team doing the research. 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[22]. In collaboration networks, edges are not directed, and the flow of knowledge is in both directions.

Papers contribute to edges according to the following 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 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[23].

To create manageable networks, we subdivide 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 the data up 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 in the year of publication. 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. One, the paper is associated with one institution-location-field: This implies an internal collaboration in the node or department. In this case an edge of weight one is added from the node to itself. Two, the number of authors is more than the number of nodes. This means that collaboration occurs within and between nodes or departments. 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

[22] 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 $\kappa = 0$.

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

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. The third case occurs when the number of authors equals the number of nodes. There is only one author per node. Hence, collaboration occurs only between nodes. 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.

As with citations, we 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

In Section II we presented a version of the linear model of social interactions applied to citation networks composed of scientists and engineers. In this framework, the demand for scientists and engineers is a linear function of citation centralities. If the centralities are positive (negative) and strategic complementarity (substitution) prevails, scientific influence increases (decreases) the demand for scientists. In this section we discuss the centralities that we use in this paper, while in Section V we use the centralities in empirical tests.

Creation of nodes and edges of the citation networks has set the stage for centralities needed to represent the influence of networks on employment of scientists in the nodes, as we showed in Section II. Specifically, we find that the in-degree centrality and the PageRank centrality have interesting properties for citation networks of scientists and engineers. All centralities are calculated by field.

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

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

In equation (7) 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 weighted by the fractions defined in Section IV.B. The in-degree centrality has the advantage of retaining non-normalized differences in scientific influence, but it has the disadvantage that it represents scientific influence only over immediate neighbors. Simple statistics on the in-degree centrality are shown in Table 1. As with the PageRank centralities to follow, the in-degree centrality is larger in nodes known to be more influential, for example Harvard Medical School and Johns Hopkins Medical school in medicine, Stanford University in computer science, and MIT in physics.

The citation PageRank centrality is defined as

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

In equation (8), 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 node j to past node 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 two advantages: first, that it measures scientific influence in the entire network, and second, that division by out-degree cancels false transmission of centrality from more central nodes to less central nodes. A disadvantage of PageRank is that the sum of the centralities by field is normalized to 1.0, limiting differences in scientific influence. Simple statistics on the PageRank centrality are also shown in Table 1.

V. Econometric Findings

Equations (5) and (6), the best response demand for research scientists and the supply of research output, are the equations to be estimated. At issue is whether the centralities increase or decrease employment of scientists, and are therefore strategic complements or substitutes. Since we find that citation and collaboration centralities are collinear (see Appendix A), we concentrate henceforth 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. Figures 1 and 2 present the relationship between research scientists and the two types of centralities. The fitted lines, for a representative sample of eight out of twenty fields, show that employment is a rising function of centrality, indicating strategic complementarity. These are visual employment-centrality coefficients.

However, endogeneity problems could bias these centrality coefficients. For instance, omitted attributes of nodes could be correlated with the centralities, resulting in a correlation between the centrality and the error term in the best response demand equation. We deal with this problem by random assignment of centralities from another node in the same field and institutional type, but from a previous time interval and hence a different network. This instrument breaks any link between omitted variables and the centralities and eliminates the bias in the centrality coefficient. Table 2 provides a discussion.

Part A. of this section describes the covariates used in the regression analysis. Part B explains 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., Tables 3 through 7. Part D. concludes with 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 7 and 8 report these results.

A. Description of the Covariates

Table 1 reports descriptive statistics for the covariates: the number of observations, mean, standard deviation, minimum, and maximum. The panel of nodes contains 77,073 observations during 1981-2011. Recall, 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 seventy-eight percent of the sample of 77,073 observations. This sample size reflects a sixty-paper cutoff during each six-year interval. Separately 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 part of Table 3 is concerned with the best response equation. Line one describes all research scientists over different tenures with a node. This includes transient researchers and researchers publishing intermittently, as well as long-term researchers. Line two describes long-term research scientists only. These are steadily publishing researchers who appear with a node for at least six consecutive years. Mean all research scientists is 261, while mean long-term research scientists is 180. Long-term research scientists are in the majority except for top departments, indicating substantial numbers of graduate students and visitors in these nodes. The maximum number of researchers is in the tens of thousands. This occurs in top medical schools. It is due to collaborations with research scientists employed as physicians in private practice or as faculty in smaller institutions. Top medical schools have many outside collaborators and our method makes the 'invisible college' for these schools visible.

Lines three and four report citation in-degree and PageRank centralities. The in-degree centrality has a mean of 1,687.54. There are about one hundred papers per node, which are cited roughly sixteen times, hence this result. The table shows how different the PageRank centrality is from the in-degree centrality. PageRank averages 0.01 because the centrality is normalized, summing to 1.0 over nodes in a field. The average number of nodes per field is one hundred, accounting for this result.

Line five contains an indicator for employment in the same field in other locations in the same institution. Universities represent one location in the data; multiple location institutions are firms, government agencies or laboratories, or nonprofit biomedical research centers. This variable could indicate substitution between locations. It occurs eighteen percent of the time.

Line six is an indicator for defense contractor firm[24]. Figure 3 shows that funding of defense contractors declines during the post-Reagan years, which predominate in our data. This suggests a

[24] Twelve defense contractors are in the data: GENERAL ELECTRIC CO, FORD MOTOR CO, GENERAL MOTORS CO, GTE CORP, ALLIEDSIGNAL INC, BOEING CO, HERCULES INC, MOTOROLA INC, RAYTHEON CO, ROCKWELL INTL CORP, TEXAS INSTRUMENTS INC, and TRW INC.

negative sign on defense contractor. However, defense contractors also tend to be large, diversified firms. This suggests a positive sign and in fact the sign of defense contractor is mixed in Tables 3 and 5. Not shown is the defense intramural dummy. It is an indicator for defense agencies and federal laboratories engaged in defense research[25]. Figure 3 shows little trend in funding of defense intramural institutions. We therefore omit this variable from the regressions.

Lines seven and eight of Table 3 complete coverage of the best response demand equation. Line seven reports average distance to top research cities. Line eight reports the number of research scientists elsewhere in the same Metropolitan Statistical Area (MSA) per million population. Average distance to leading research cities is 1,163 miles[26]. Learning from innovative research could be more costly at greater distances. Hiring could occur less often. These factors could weaken the influence of high impact research as distance increases.

Average research scientists in the same field elsewhere in an MSA is divided by MSA population in millions. Its mean value is 234. This variable probably stands for competition among research institutions. And indeed, its sign is negative in the regressions.

Table 3 concludes with variables entering the research output equation: the knowledge production function. The dependent variable is the number of citations received over the next five years, which is not available after 2006 since the data end in 2011. The sample attrition which this causes shows up in a smaller sample size in this equation. The main independent variables, following equation (6), 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. At means, the squared value of all research scientists is 519,115 and the squared value of long-term research scientists is 312,950.

[25] 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.

[26] The top seven research cities are Boston, New York, Philadelphia, Washington-Baltimore, Chicago, San Francisco, and Los Angeles.

B. Discussion of Centrality Instruments

Table 2 discusses instrumental variables for the centralities. For each centrality (in-degree centrality, PageRank centrality), the instrument is lagged by one or two six-year intervals. The instrumented centrality is randomly assigned to a node from another node in the same field and institutional type, or else it is a randomly assigned 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 centrality instruments in Table 2: same field versus most cited other field, and one-period lag versus two-period lag. 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 previously described. 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[27]. We draw a uniform random sample of one observation for each citing institution and field to define the instrument for the observed centrality[28]. Resulting instruments for the centralities exhibit a much lower correlation with research scientists on the left of the best response demand equation: see (5) in Section II. For example, the correlation between the in-degree centrality and all research scientists is 0.88, but the correlation between the randomly assigned instrument for the in-

[27] 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 paper, the same field of science. Previous code removes same institution from the sample, ruling out self-citation.

[28] The instrument for the centrality is different from the mean centrality for a field because it would not be identified once field fixed effects are included, as they are in the 2SLS and 3SLS results reported below.

degree centrality and all research scientists is 0.26. This suggests that we have removed most of the correlation between error term and the centrality.

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. Equation (5) would imply that both citation and collaboration centralities belong in the best response demand equations. However, the two centralities are highly correlated, and their effects cannot be distinguished[29]. Hereafter, we confine the estimation to the citation centralities, though for transparency we report expanded regressions including both centralities in the online Appendix A.

D. Estimates of Best Response Demand Equations

Table 3 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 of the table and long-term research scientists on the right. All equations include fixed effects for institution, field, and six-year period. Table 3 assumes robust errors or else errors clustered by node. Randomly assigned instruments for centralities were taken from the network one period before. For example, the 2006-2011 network drew instruments from the 2001-2006 network.

Equations 3.1 and 3.4 are the benchmark OLS results. The coefficient of the citation in-degree centrality is positive and significant. It is larger for all research scientists (0.137) than for long-term research scientists (0.084). This is because high centrality nodes employ a larger share of short-term research scientists. Equations 3.2-3.3 and 3.5-3.6 are 2SLS results where robust or clustered standard errors are assumed. The coefficient on in-degree centrality stays about the same, but as expected 2SLS standard errors increase substantially.

[29] 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.

The indicator for other locations has a negative sign. Multiple location nodes employ fewer scientists, perhaps due to substitution between locations. Defense contractor dummy is negative but not always significant.

Average distance to leading research cities has a negative effect on employment of research scientists. The negative coefficient of research scientists in the same field elsewhere in the same MSA per million population suggests crowding out of research by other local institutions.

We conclude Table 3 with instrument tests. First, excluded instruments for the centralities are highly significant in the first stage equations. Second, the robust chi-square tests for endogeneity are on the margin of significance, though they are highly significant in Table 5 using the PageRank centrality. Third, tests of the overidentifying restrictions accept the null hypothesis that the excluded instruments are valid. These tests are available under the robust VCE assumption of equations 3.2 and 3.5, but not for the clustered errors assumption of 3.3 and 3.6.

Table 4 explores differences by type of institution in the coefficient of in-degree centrality. The method is 2SLS. The table interacts dummies for type of non-university institution with the centrality and enters the interaction as a separate endogenous variable. The excluded type is universities. In equations 4.1 and 4.3 the non-university types are defense institutions (defense contractor firms and defense government institutions), non-defense government institutions, firms, and biomedical research centers. Equations 4.2 and 4.4 break up defense institutions into defense government institutions and defense contractor firms. All else is the same.

Since the specification follows 3.2 of Table 3 except for the added interaction terms, Table 4 concentrates on the citation in-degree centrality and its interactions. For instruments we use the centrality instruments and their interactions with the dummies for institution type. The primary finding is that universities exhibit coefficients on the in-degree centrality that are at least as large as the coefficients for other types. Strategic complementarity is therefore greater for universities. This is especially true of defense contractors, where the sum of the main effect and the interaction with the defense contractor type is less than zero, suggesting strategic substitution rather than strategic complementarity, the sole example

of this in our data. The strategic complementarity is less for non-defense firms and biomedical research centers than for universities, but it is still greater than zero.

Table 5 replaces in-degree centrality with PageRank centrality. All research scientists are the dependent variable in 5.1-5.3, while long-term research scientists are the dependent variable in 5.4-5.6. Equations 5.1 and 5.4 contain the benchmark OLS results. 2SLS results are found in 5.2-5.3 and 5.5-5.6. 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 other nodes using the research of a given node stimulates employment of research scientists in the node in question. This again defines strategic complementarity.

In Table 6 the lag on the centrality instruments is two six-year periods, unlike Tables 3-5, where it is one period. For example, the 2006-2011 network draws on the 1996-2001 network for its randomly assigned centrality instruments. Throughout the method is 2SLS. The table concentrates on centrality findings since findings on the other covariates are similar to those in the previous tables. Panel A reports findings using the citation in-degree centrality, while Panel B report findings for citation PageRank.

In Panel A the coefficient on citation in-degree centrality is positive, highly significant, and about the same as in Table 3. In Panel B the coefficient on the PageRank centrality is positive, significant, but nearly twice the coefficient in Table 5. The increase in the coefficient suggests that the two-period PageRank instrument is superior to its one-period counterpart. The evidence of Table 6, as before, strongly supports the hypothesis of strategic complementarity in networks of scientists and engineers. 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 (5) and (6), which are reproduced below for convenience:

$$(5) \qquad \ell_{Ri} = \frac{1}{\eta_R}[x'_{Ri}\theta_R + \gamma_R C_{Ci} + \gamma_T C_{Ti} + u_{Ri}]$$

$$(6) \qquad q_{Ri} = v'_{Ri}\varphi_R + \frac{\eta_R}{2}\ell^2_{Ri} + e_{Ri}$$

Notice that (6) requires a measure of research output q_{Ri} and an instrument for research scientists squared (ℓ^2_{Ri}), since scientists are correlated with the error term e_{Ri}.

We use fractional citations received by a node over the next five years to measure research output. By fractional citations we mean citations adjusted by the share of authors in a node out of all authors listed on papers associated with the node. Requiring more years than this would result in increased sample attrition. As it is, taking account of citations over the next five years eliminates years 2007-2011, since the citations are not fully observed after 2006.

An instrument is needed for research scientists squared on the right of (6). An excellent choice would be an exogenous counterpart to past federal R&D funding. That should be highly correlated with research scientists because it supports past graduate education. It could be a compelling instrument, because federal money is the most important source of support for research scientists in universities. We know that federal support of R&D in universities varies among fields and that it is concentrated among top university departments. Finally, we know that research scientists working outside universities receive federal support through faculty grants in universities during their graduate school training. These pieces of information factor into the funding instrument described below.

We compute the instrument as follows. For universities, which form two-thirds of our observations, we calculate average federal R&D funding in millions of 2012 dollars during the first three years of the data series for each university and field. The data come from the National Science Foundation Survey of Higher Education R&D (HERD), which begins in 1973. For many fields, the first three years

are 1973-1975, long before the investigation begins. In detailed engineering and environmental sciences, the first three years are 1980-1982 (materials science is the lone 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 at least the period 1981-1986, and sometimes the period 1986-1991, the first three years of funding precede the analysis by at least 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 acts 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 government policy by imposing the national growth rate of federal funding in a field.

In the next few paragraphs, we express these ideas 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:

$$\overline{RD}_{ij} = \sum_{t=1}^{3} RD_{ijt}\,/3 \tag{9}$$

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

$$RD_{ijt}^{Sim} = \overline{RD}_{ij}\left(1 + g_{32}^{i}\right) \cdot \left(1 + g_{43}^{i}\right) \cdot \ldots \cdot \left(1 + g_{t,t-1}^{i}\right) \tag{10}$$

In equation (10) $g_{t,t-1}^{i} = \left(RD_{t}^{i} - RD_{t-1}^{i}\right)/RD_{t-1}^{i}$. 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 (10) interact the initial three-year average $\overline{RD}_{ij}$ for a 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:

$$RDK_{ijt}^{Sim} = \sum_{l=1}^{8}(1-d)^{l-1} RD_{ij,t-l}^{Sim} \tag{11}$$

The stock in (11) is depreciated by a depreciation rate of $d = 0.15$. Equation (11) 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 (9) leading to the stock (11)[30]. 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:

$$\overline{RDK}_{it} = \sum_{l=1}^{8}(1-d)^{l-1}\overline{RD}_{i,t-l} \qquad (12)$$

Equations (11) and (12) constitute the funding instrument for research scientists squared in (6), while the instrument for the centrality in (5) is as explained in Section V.B.

Table 7 reports 3SLS results with clustered errors, using the citation in-degree centrality in the best response demand equation for research scientists. For conciseness we limit reporting to the variables of interest, which are the centralities and research scientists squared. The centrality coefficients are the same (System 7.1) or higher (System 7.2) compared to the two-stage least squares results of Table 3. They remain positive and highly significant. Therefore, strategic complementarity also prevails in these two-equation 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 research output[31]. At means, the square of all research scientists in Table 7, System 7.1 is 519,115.1, which corresponds to $0.000308 \times 519{,}115.1 = 159.1606$ five-year citations produced. At the 95th percentile the square of all

[30] 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.

[31] 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 0.000308 for all research scientists in Table 7, System 7.1, and recalling from Table 3 that the mean of all research scientists is 260.60, the derivative is $2 \times 0.000308 \times 260.68 = 0.1606$. At means, an additional research scientist delivers 0.1606 citations over the next five years. The derivative for System 7.2, evaluated at a mean of 179.68 for long-term researchers, is $2 \times 0.0005153 \times 179.68 = 0.1852$. However, these results are evaluated at means, whereas the payoff is much greater in the upper ranges of employment of research scientists in large research groups. For example, at the 95th percentile the number of all research scientists is 972, and the derivative is $2 \times 0.000308 \times 972 = 0.5988$. Likewise, the 95th percentile number of long-term researchers is $2 \times 0.0005153 \times 753 = 0.7768$. The payoff is largest at the top of the distribution.

researchers is 944,784 and predicted five-year citations are $0.000308 \times 944{,}784 = 290.9935$. At the 99th percentile, squared researchers are 8,643,600 and predicted citations rise to $0.000308 \times 8{,}463{,}600 = 2,606.7888$. Compared to mean employment, the 99th percent level of employment drives five-year citations that are over sixteen times greater (2606.7888/519,115.1=16.3784). The results for the citations equation of System 7.2 are quite similar.

Table 8 replaces the citation in-degree centrality with the citation PageRank centrality. The centrality coefficient estimates remain positive and highly significant, but they are much higher than in the two-stage least squares results of Table 5, perhaps because 3SLS takes cross-equation errors into account. This concludes the empirical work, which strongly supports the hypothesis of strategic complementarity between nodes of the networks, and which implies a rising supply curve of research output in the final tables, Tables 7 and 8, especially in the largest (and top) nodes of the networks.

VI. 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, which cover all major U.S. research institutions and fields of science, we find that the networks have grown more dense, less fragmented with time, and 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. Since intramural defense scientific research grows slowly with time, the primary trend is a decline in defense-supported scientific research in defense contractor firms following the breakup of the Soviet Union. This factor has the opposite effect on network connections from the expansion of federal support of academic research.

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, government 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 government policy toward that field? This question, which is related to issues raised by Milton Friedman, is not simple to answer, because centralities depend on the level of funding, requiring a more primitive concept of centrality[32]. 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

[32] See Friedman (1962) and Friedman (1981).

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.

_____, Ameya M. Haté, and J. Roger Clemmons. (2026). "Construction of Networks of Scientists and Engineers in the United States, 1981-2011." Working Paper, Rensselaer Polytechnic Institute.

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.

Friedman, Milton (1962). "The Role of Government in Education," Chapter 7 of **Capitalism and Freedom**. Chicago: University of Chicago Press.

_____ (1981). "An Open Letter on Grants." **Newsweek** 18 May 1981, p. 99.

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
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	77,073	260.68	809.01	0	36,973
Long-Term Research scientists	77,073	179.68	548.79	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 Locations (1 if Yes, 0 if No)	77,073	0.183	0.387	0	1
Defense Contractor Dummy	77,073	0.017	0.128	0	1
Average Distance to Top Research Cities	77,073	1,162.94	320.19	852.24	1,847.02
Research Scientists, Same Field, Rest of MSA, per million MSA Population	77,073	234.29	779.34	0	24,661.30
Research Output Equation [b]					
Citations Received, Next 5 Years	64,447	490.55	1,808.16	0	80,018.61
All Research Scientists Squared, Lagged Two Years	58,415	519,115.1	8,900,253	0	8.52e+08
Long-Term Research Scientists Squared, Lagged Two Years	58,415	312,950.4	5,403,460	0	4.71e+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 2
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 3
Best Response Demand for Research Scientists
Citation In-Degree Centrality
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(3.1)	(3.2)	(3.3)	(3.4)	(3.5)	(3.6)
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**	0.147**	0.147**	0.084**	0.079**	0.079**
	(0.002)	(0.006)	(0.011)	(0.003)	(0.006)	(0.010)
Same Institution: Other Locations, Same Field (1 if Yes, 0 if No)	-54.15**	-32.27*	-32.27	-53.75**	-65.33**	-65.33**
	(6.19)	(12.78)	(28.82)	(6.56)	(13.45)	(25.59)
Defense Contractor Dummy (1 if yes, 0 if no)	-24.26*	-38.22**	-38.22	-29.16**	-21.83	-21.83
	(12.49)	(13.07)	(21.82)	(10.95)	(13.58)	(20.46)
Average Distance to Top Research Cities	-0.042**	-0.028*	-0.028	-0.052**	-0.059**	-0.059
	(0.010)	(0.012)	(0.037)	(0.009)	(0.011)	(0.034)
Research Scientists, Same Field, Rest of MSA, per million MSA Population	-0.051**	-0.054**	-0.054**	-0.029**	-0.027**	-0.027
	(0.004)	(0.005)	(0.015)	(0.004)	(0.004)	(0.015)
F-Test for Excluded Instruments		32.895	8.690		32.847	8.693
[P-Value]		[0.0000]	[0.0002]		[0.0000]	[0.0002]
Robust Chi-Square (1) Test for Endogeneity [P-Value]		3.252			0.837	
		[0.07]			[0.36]	
Test of Overidentifying Restrictions		0.228			2.663	
[P-Value]		[0.632]			[0.103]	
Number of Observations	51,052	51,052	51,052	51,052	51,052	51,052
F (284, 50,892)	502.91			336.31		
Probability>F	0.0000			0.0000		
Wald Chi-Square		138,806.93	164,162.23		93,260.90	118,411.82
[P-Value]		[0.0000]	[0.0000]		[0.0000]	[0.0000]

Table 3
Best Response Demand for Research Scientists
Citation In-Degree Centrality
OLS and 2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(3.1)	(3.2)	(3.3)	(3.4)	(3.5)	(3.6)
R-squared	0.926	0.923	0.923	0.842	0.841	0.841
Root MSE	262.91	267.51	267.51	260.53	261.33	261.33

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 4
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	(4.1)	(4.2)	(4.3)	(4.4)
Dependent Variable	All Research Scientists		Long-Term Research Scientists	
Method	2SLS			
VCE Assumption	Robust			
Lag on Instruments for Citation In-Degree Centrality	1 Period			
Citation In-Degree Centrality	0.149**	0.149**	0.079**	0.079**
	(0.006)	(0.006)	(0.007)	(0.007)
Citation In-Degree Centrality × Defense Institution	0.011		0.011	
	(0.012)		(0.011)	
Citation In-Degree Centrality × Defense Government Institution		0.012	-0.009	0.013
		(0.012)	(0.007)	(0.011)
Citation In-Degree Centrality × Nondefense Government Institution	-0.017*	-0.017*		-0.009
	(0.009)	(0.009)		(0.007)
Citation In-Degree Centrality × Defense Contractor Firm		-0.331		-0.698**
		(0.185)		(0.226)
Citation In-Degree Centrality × Nondefense Firm	-0.084*	-0.087*	-0.149**	-0.153**
	(0.043)	(0.043)	(0.044)	(0.045)
Citation In-Degree Centrality × Biomedical Research Center	-0.048**	-0.048**	-0.031**	-0.031**
	(0.007)	(0.007)	(0.007)	(0.007)
Number of Observations	51,052	51,052	51,052	51,052
Wald Chi-Square	176,435.94	177,295.31	100,776.88	96,595.27
[P-Value]	0.0000	0.0000	0.0000	0.0000
R-squared	0.923	0.923	0.835	0.835
Root MSE	268.56	268.51	265.60	265.93

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 5
Best Response Demand for Research Scientists
Citation PageRank Centrality
2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(5.1)	(5.2)	(5.3)	(5.4)
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 PageRank Centrality	1 Period	1 Period	1 Period	1 Period
Citation PageRank Centrality	28,044.47**	28,044.47*	17,242.93**	17,242.93**
	(4,781.04)	(13,017.22)	(3,169.23)	(8,269.84)
Same Institution: Other Locations, Same Field (1 if Yes, 0 if No)	-398.08**	-398.08**	-265.64**	-265.64**
	(18.64)	(73.94)	(12.46)	(49.79)
Defense Contractor Dummy (1 if yes, 0 if no)	125.82**	125.82**	62.59**	62.59*
	(32.52)	(48.63)	(22.63)	(31.81)
Average Distance, Top Research Cities	-0.054	-0.054	-0.055*	-0.055
	(0.042)	(0.149)	(0.028)	(0.097)
Research Scientists, Same Field, Rest of MSA, per million MSA Population	-0.005	-0.005	0.005	0.005
	(0.008)	(0.032)	(0.007)	(0.031)
F-Test for Excluded Instruments	25.738	3.429	25.734	3.441
[P-Value]	[0.0000]	[0.0322]	[0.0000]	[0.0322]
Robust Chi-Square (1) Test for Endogeneity [P-Value]	9.782		5.015	
	[0.0018]		[0.0251]	
Test of Overidentifying Restrictions	0.186		0.130	
[P-Value]	[0.6661]		[0.7189]	
Number of Observations	51,052	51,052	51,052	51,052

Table 5
Best Response Demand for Research Scientists
Citation PageRank Centrality
2SLS Estimates, One Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(5.1)	(5.2)	(5.3)	(5.4)
Wald Chi-Square	40,690.68	53,285.45	50,451.35	15,225.13
[P-Value]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
R-squared	0.406	0.406	0.420	0.420
Root MSE	744.95	744.95	498.14	498.14

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
Citation In-Degree Centrality
2SLS Estimates, Two Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(6.1)	(6.2)	(6.3)	(6.4)
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
Panel A. Citation In-Degree Results				
Citation In-Degree Centrality	0.138**	0.138**	0.079**	0.079**
	(0.005)	(0.010)	(0.005)	(0.010)
F-Test for Excluded Instruments	39.68	6.404	39.547	39.547
[P-Value]	[0.0000]	[0.0017]	[0.0000]	[0.0000]
Robust Chi-Square (1) Test for Endogeneity [P-Value]	0.084 [0.7717]		0.477 [0.4896]	
Test of Overidentifying Restrictions [P-Value]	1.004 [0.3163]		4.281 [0.0385]	
Number of Observations	42,045	42,045	42,045	42,045
Wald Chi-Square	142,458.83	479,051.90	98,977.94	98,977.94
[P-Value]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
R-squared	0.930	0.930	0.843	0.843
Root MSE	273.80	273.80	274.34	274.34

Table 6
Best Response Demand for Research Scientists
Citation In-Degree Centrality
2SLS Estimates, Two Period Lag
(Standard Errors in Parentheses)

Variable or Statistic	(6.1)	(6.2)	(6.3)	(6.4)
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
Panel B. Citation PageRank Results				
Citation PageRank Centrality	37,711.44**	37,711.44*	22,594.96**	22,594.96*
	(6,840.41)	(17,849.73)	(4,246.91)	(11,030.34)
F-Test for Excluded Instruments	19.63	3.41	19.62	3.41
[P-Value]	[0.0000]	[0.0331]	[0.0000]	[0.0332]
Robust Chi-Square (1) Test for Endogeneity [P-Value]	16.252 [0.0001]		0.084 [0.7717]	
Test of Overidentifying Restrictions [P-Value]	3.102 [0.078]		1.004 [0.3163]	
Number of Observations	42,045	42,045	42,045	42,045
Wald Chi-Square	62,565.91	100,988.61	67,004.27	28,428.94
[P-Value]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
R-squared	0.397	0.397	0.413	0.413
Root MSE	802.30	802.30	529.95	529.95

Notes. Data are based on a panel of nodes, which are defined as institutions, fields, and locations. Data covers the years from 1986-2011. 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 7
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 7.1 All Research Scientists		System 7.2 Long-Term Research Scientists	
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 Errors		Clustered Errors	
Lag on Instruments for Citation Centrality	1 Period		1 Period	
Funding Instruments for Research Scientists	Lagged Stocks of Federally Funded R&D in Millions of 2012 $		Lagged Stocks of Federally Funded R&D in Millions of 2012 $	
Citation In-Degree Centrality	0.1371** (0.0052)			
All Research Scientists Squared, Lagged Two Years		0.0003080** (0.0000878)		
Citation In-Degree Centrality			0.1060** (0.0043)	
Long-Term Research Scientists Squared, Lagged One Year				0.0005153** (0.0001495)
Number of Observations	37,336	37,336	37,336	37,336
Chi-Square	344,239.49	4,591.86	165,818.64	5,533.93
[P-Value]	0.0000	0.0000	0.0000	0.0000

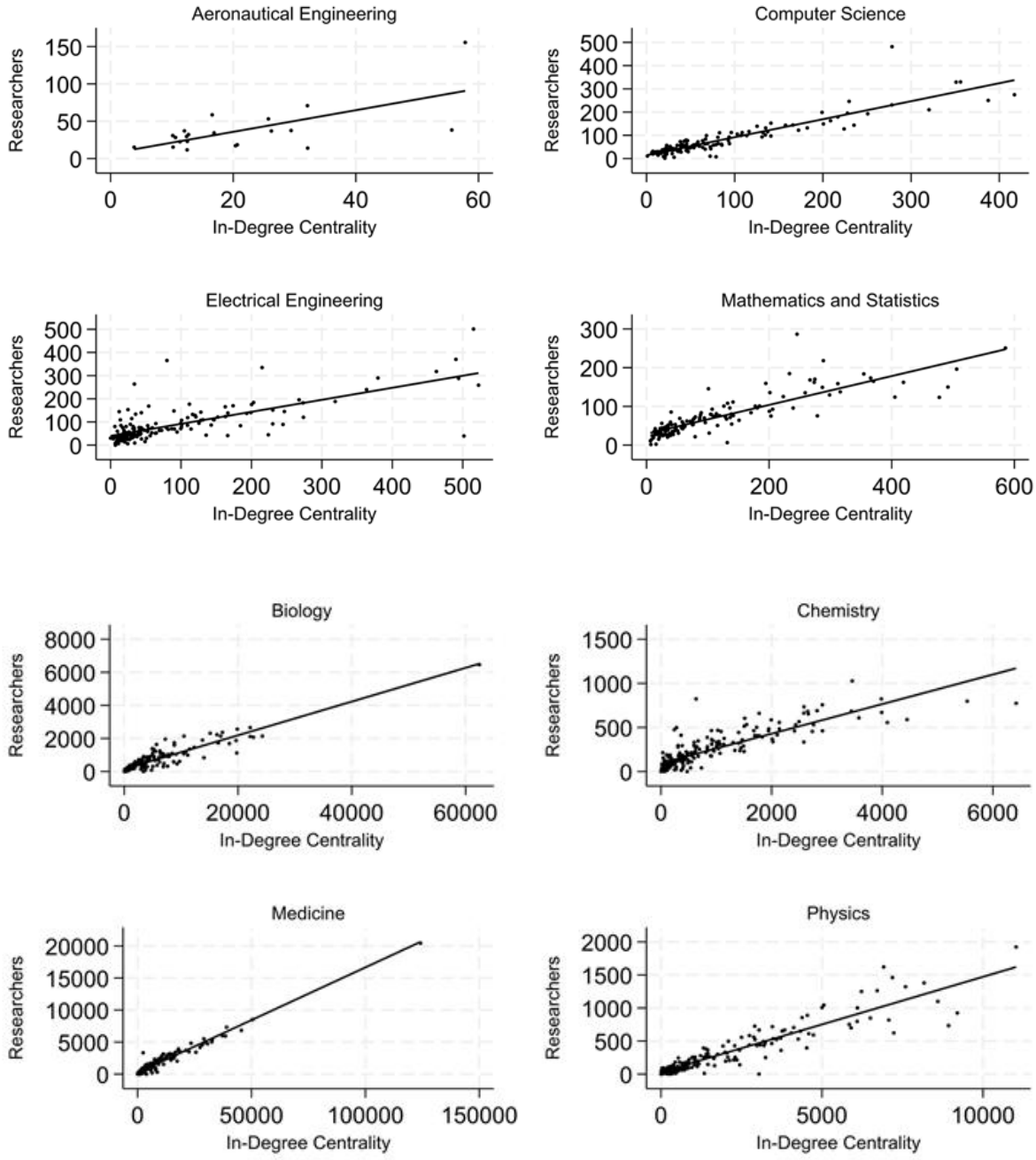
Notes. Data are based on a panel of nodes, which are defined as institutions, locations, and fields. Data covers the years from 1986-2006. Period 1981-1986 is deleted since IVs are lagged one period. Period 2007-2011 is deleted since citations are incomplete. See the text and Table 1 for a description of the covariates. [a] Dependent variable in Eq. 1 is All Research Scientists (System 7.11) or Long-Term Research Scientists (System 7.2). [b] In both System 7.1 and System 2 the dependent variable in Eq. 2 is total citations received over the next five years. Eq. 1 follows 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 8
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 8.1 All Research Scientists		System 8.2 Long-Term Research Scientists	
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	Lagged Stocks of Federally Funded R&D in Millions of 2012 $		Lagged Stocks of Federally Funded R&D in Millions of 2012 $	
PageRank Citation Centrality	161,589.10** (46,043.91)			
All Research Scientists Squared, Lagged One Year		0.0003199** (0.0000898)		
Citation PageRank Centrality			125,221.70** (35,673.79)	
Long-Term Research Scientists Squared, Lagged One Year				0.000536** (0.0001530)
Number of Observations	37,336	37,336	37,336	37,336
Chi-Square	28,569.28	9,606.82	1.61e+06	7,079.42
[P-Value]	0.0000	0.0000	0.0000	0.0000

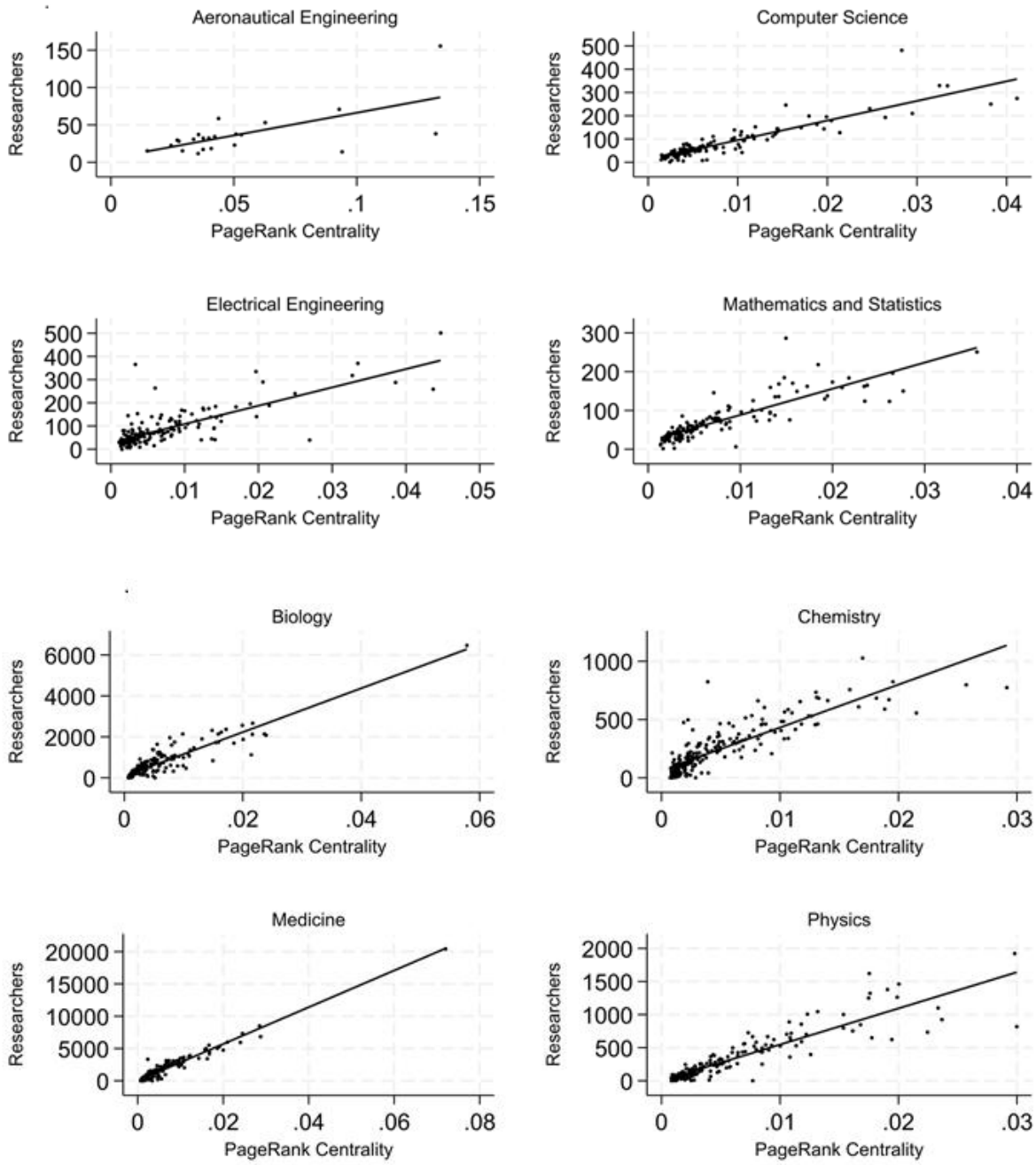
Notes. Data are based on a panel of nodes, which are defined as institutions, locations, and fields. Data covers the years from 1986-2011. Period 1981-1986 is deleted since IVs are lagged one period. Period 2007-2011 is deleted since citations are incomplete. See the text and Table 3 for a description of the covariates. [a] Dependent variable in Eq. 1 is All Research Scientists (System 8.1) or Long-Term Research Scientists (System 8.2). [b] In both System 8.1 and System 8.2 the dependent variable in Eq. 2 is total citations received over the next five years. Eq. 1 follows 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
Researchers by In-Degree Centrality and Field



Source: Web of Science and authors' calculations

Figure 2
Researchers by PageRank Centrality and Field



Source: Web of Science and authors' calculations

Figure 3
Federally Funded Defense R & D, 1981-2011



Source: National Science Foundation

On-Line Appendices

Appendix A

Table A1
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	(A1.1)	(A1.2)	(A1.3)	(A1.4)	(A1.5)	(A1.6)
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 B

Table B1
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 B1
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 B2
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.