Peer effects in cancer research

Job Market Paper

Adhen Benlahlou

November 29, 2024

This paper develops a theory of scientific peer effects to study scientific productivity when authors care about the behavior of their coauthors. The theory predicts that an author's output increases in their Katz-Bonacich centrality, a standard measure of centrality in networks. I test the model's predictions using a detailed dataset of cancer researchers, showing that an author's position in the coauthorship network (measured by Katz-Bonacich centrality) is a key determinant of their productivity. After controlling for observable characteristics and unobservable network-specific factors, my results demonstrate the importance of peer effects not only from superstar scientists but from a broader range of collaborators. These findings highlight the relevance of network externalities in scientific productivity and provide empirical evidence for the critical role of network structure in fostering knowledge creation. By controlling for prestige-driven effects, my results suggest that peer effects are driven by scientific complementarities rather than status. Additionally, the Nash-Katz-Bonacich linkage offers strong policy implications, particularly for designing optimal network structures to maximize collective knowledge production.

1 Introduction

Knowledge and technology are crucial drivers of economic growth [Romer, 1990, Weitzman, 1998], and the production and diffusion of knowledge have become a focal point of research. Recent studies indicate a shift in the structure of knowledge production, with collaboration on the rise across nearly all scientific fields [Wuchty et al., 2007, Jones, 2009, Agarwal et al., 2015]. Scientists are increasingly interconnected through extensive collaboration networks. This study explores how a researcher's work is influenced by other scientists within their network, which is termed by the literature 'scientific peer-effects'. A fundamental inquiry in economics of science is how to link position in the coauthorship network and research productivity.

This paper is the first to quantify heterogeneity in scientific peer effects by examining the combined influence of network position and individual characteristics on the productivity of scientists. The microfoundations of this study rest on two fundamental premises. First, author outcomes can be divided into an idiosyncratic component and a peer effect component; a scientist's productivity is therefore influenced both by their

individual characteristics and by the influence of their peers. Second, at the group level, peer effects encompass the collective influence that group members exert on each other, including the bilateral cross-influences within the group. This model formally links a scientist's productivity to her network position captured by a centrality measure defined by Bonacich [1987].

The underlying theoretical framework establishes the behavioral basis for the estimation of a variation of the high-order spatial autoregressive (SAR) model [Liu and Lee, 2010]. Within this framework, the main parameter of interest captures the strength of the dyadic influences within the network. This parameter plays a crucial role in the computation of the Katz-Bonacich centralities, as it serves as the decaying weight for path length. However, estimating this model poses certain challenges.

The challenge in determining the presence of peer effects comes from a need to simultaneously consider the two key elements at play in coauthorship synergies (strategic complementarities) between coauthors partners and the matching process bringing these researcher together. First, at the coauthorship level, higher levels of research output are associated with increased effort and collaboration between coauthors. Given that each partner brings a unique bundle of attributes to the collaboration, it is crucial to consider the multi-dimensional nature of these interactions in order to capture the full scope of synergies within coauthorship networks. Second, coauthorships are voluntary collaborations formed around mutual interests, and as such, the selection of partners is not random. Importantly, a scientist's decision to engage in co-authorship, and thereby their position within the broader network, is often the result of a strategic choice. This decision is typically influenced by expectations regarding the potential outcomes of the collaboration. Consequently, coauthorship networks are endogenously determined, meaning they are shaped by strategic choices made by researchers within them. This complexity is further compounded by the presence of unobservable factors that affect both a researcher's productivity and their propensity to collaborate [Mairesse and Turner, 2006, Fafchamps et al., 2010, Freeman et al., 2014].

Tackling this complexity with existing methodological tools is challenging. Prior work has exploited the unexpected deaths of scientists as an exogenous shock to assess the impact of coauthorship on productivity [Azoulay et al., 2010, 2019, Jaravel et al., 2018, Oettl, 2012]. In this paper, I exploit an instrumental variable approach proposed by Lee et al. [2021], using a logistic regression to predict link formation probabilities between all pairs of scientists and using the predicted network as an instrument. These predictions are based on exogenous dyadic characteristics such as whether the individuals belong to the same institution, share similar demographics, and other relevant characteristics. This approach is robust even in networks where connections are endogenously determined and simplifies the computational complexity by bypassing the need to model the entire network formation process.

I test the proposed peer effects model on the coauthorship network of cancer researchers. For this purpose, I compiled a dataset that combines CVs gathered from university websites and professional networks, disambiguated publication data from the PubMed Knowledge Graph [Xu et al., 2020], and additional metadata from the Torvik research group. Cancer research is particularly relevant as it relies heavily on collabora-

tion and tacit knowledge, which is best transferred through direct interaction [Polanyi, 1966, Nonaka and Takeuchi, 2007]. Knowledge production in this field is often collective, and output is measured by the number and quality of publications [Katz and Martin, 1997]. Additionally, cancer research drives significant innovation and has a major economic impact; a 1% annual reduction in mortality from cancers such as lung, colorectal, breast, leukemia, pancreatic, and brain cancer could lower productivity costs by $814 million per year [Bradley et al., 2008].

The first result of this study is consistent with the theory. Scientists' Bonacich centralities have highly significant effects on their scientific production. In particular, the aggregate equilibrium effort and hence total output increases with connectivity. Overall, in equilibrium, researchers with the highest centrality are also those with the highest production. This result is robust to many natural controls suggested by the existing literature, including seniority, measures of past production, prestige, and gender. These results are illustrative, indicating the potential significance and magnitude of network externalities in scientific productivity.

Second important result, among the most commonly used centrality measures, Bonacich centralities distinguish themselves from other network measures by their dyadic, microfoundation, and global features. I compare the explanatory power of the theory-driven centrality measure with standard centrality measures, and the results support the proposed microfoundation. This sheds light on the mechanisms underlying peer effects. The precise description of Nash equilibrium behavior through a network measure is particularly beneficial. For example, the Nash-Katz-Bonacich linkage carries significant implications for both comparative statics and the design of optimal network policies [Ballester et al., 2006].

This study contributes to research on peer effects in science by examining both direct and indirect impacts of coauthorship, alongside individual characteristics. While previous studies have focused on immediate, localized peer effects—often treating them as homogeneous [Agrawal et al., 2017, Azoulay et al., 2010, 2019, Borjas and Doran, 2015, Waldinger, 2012, 2016]—this research broadens the scope, exploring not just coauthors' direct influence but also the indirect effects across the wider scientific network. For example, Azoulay et al. [2010] find significant knowledge transfer among medical researchers, while Waldinger [2012] finds limited spillovers within university departments. Further, Waldinger [2012] observes a drop in research output after the dismissal of a coauthor, while Azoulay et al. [2010] links the death of a leading scientist to a 5-8% decline in publication quality of their coauthors. Oettl [2012] highlights that the death of prominent, helpful researchers reduces coauthors' output quality, but not quantity. Despite these findings, much of the literature focuses on egocentric, one-degree coauthorship networks, overlooking broader scientific networks. Moreover, peer effects are shown to vary by expertise Agrawal et al. [2017], gender Ding et al. [2010], helpfulness Oettl [2012], and other factors. In contrast to the typical focus on "star" scientists, my analysis examines the productivity of a broader set of cancer researchers, showing that peer effects extend beyond 'star' researchers.

From an econometric perspective, recently there has been significant progress in the literature on identification and estimation of social network models (see Bramoullé et al. [2020], for a recent survey). In applied research, the linear in means model is among the most popular models. Bramoullé et al. [2009] provide identification conditions for this model based on the intransitivities in the network structure and propose an IV-based estimation strategy exploiting exogenous characteristics of indirect connections. Yet, the validity of the IVs relies on the assumption that the network structure captured by the adjacency matrix is exogenous. If the adjacency matrix depends on some unobserved variables that are correlated with the error term of the spatial autoregressive model, then the adjacency matrix is endogenous and this IV-based estimator would be inconsistent. In this paper, by focusing on one particular field, the analysis is more tractable, and comparisons between different researchers are more valid, as different subfields of the medical science universe may have different research production processes. My choice thereby mitigates potential confounding accross subfields. To further reduce this potential bias, I use the predicted adjacency matrix based on predetermined dyadic characteristics (instead of the observed adjacency matrix) to construct IVs for this model. This allows us to estimate the causal impact of coauthorship connections.
The paper proceeds as follows. Section 2 introduces the game-theoretical framework, which models scientific production within a network accounting for knowledge spillovers. In section 3, I outline the econometric framework and discuss the empirical strategy of the paper. Section 4, details the construction of the dataset. Section 5 presents the empirical findings of the analysis, while Section 6 offers concluding remarks.

2 Analytical Framework

In this section, I first outline the conceptual framework that forms the foundation of the analysis. This outline provides the necessary theoretical background and intuition for the subsequent formal model, which I introduce in the following subsection.

2.1 Theory

At the core of the theoretical framework on scientific peer effects is the concept that scientific success is inherently collaborative, depending on the researchers' ability to work together. This collaboration manifests in various forms, including the exchange of tacit knowledge. Especially when knowledge is novel, it tends to remain tacit and is often shared within closely-knit groups. Since these interactions cannot be formalized into official contracts, they rely heavily on reciprocal trust. Scientific collaborations are crucial in building this trust, facilitating the exchange of information and knowledge, and thereby enabling the generation of new ideas.
Knowledge dissemination occurs through coauthorship links within the network, with scientists gaining knowledge not only from direct coauthors but also from indirect connections to authors further removed within the network. Thus, scientists with more connections are better positioned to acquire valuable knowl-

edge and leverage the skills of their coauthors for their projects, leading to increased productivity. My aim is to demonstrate that coauthorship connections are positively correlated with heightened scientific productivity. These considerations underpin my initial hypothesis:

$H1$: Authors who are more socially connected have higher scientific productivity.

Additionally, scientists collaborating with creative and prolific peers can access more valuable ideas. By maintaining close connections with highly productive scientists, researchers gain early access to cutting-edge ideas and emerging trends within their field. This early access is particularly crucial in the realm of scientific publication, where the novelty and timeliness of findings often dictate the impact and recognition of the research.

The dynamics of scientific collaboration mean that those who are part of a network with highly productive individuals can leverage their peers' insights, methodologies, and discoveries, effectively amplifying their own research capabilities. Engaging with productive peers allows scientists to stay at the forefront of their discipline, and to be continuously informed of the latest advancements and approaches.

Early access to valuable ideas and new methodologies directly contributes to a scientist's own productivity. It allows them to build on the latest research, refine their hypotheses, and design more effective experiments. Consequently, all else being equal, a scientist who is closely associated with highly productive peers will, on average, be more productive herself. This leads to the following hypothesis:

$H2$: Authors linked to more productive coauthors have higher productivity.

These hypotheses provide insight into the factors influencing an author's productivity. Consequently, an author's productivity can be attributed to both their individual traits and the productivity levels of their coauthors. It is crucial to consider the recursive nature of productivity—where one's productivity depends on the productivity of others—when quantifying peer effects.

There are two primary challenges in testing these hypotheses, and resolving these challenges lies at the core of my contribution. The first challenge is measuring scientific connections, as these relationships are often not directly observable by the econometrician or are only partially observable. The second challenge involves identifying the specific characteristics of scientific connections that are relevant to scientific productivity.

I begin by addressing the second inquiry. Numerous aspects of a network could potentially influence the productivity of scientists, as evidenced by the extensive literature on social networks Wasserman and Faust [1994]. It's important to clarify that my focus is not solely on social networks per se; rather, it's on investigating how these networks influence authors' productivity, particularly their ability to produce impactful and innovative scientific work.

One centrality measure that consistently emerges as crucial for understanding how social connections shape network members' behavior is the centrality measure pioneered by Katz [1953] and later refined by Bonacich

[1987]. The weighted Bonacich centrality measure expresses centrality in a recursive manner, taking into account both an author's characteristics and a weighted summation of the centralities of their connected peers. The weights are determined by the configuration of their coauthorship relationships.
To elaborate, for a given network g, we associate its adjacency matrix $G = [g_{ij}]$, which records the direct connections within g. The entry in the (i, j) cell of G^k indicates the number of paths of length k in g between nodes i and j. It is crucial to recognize that these paths may not represent only the shortest path between the authors. The $(n \times 1)$ vector of ones is denoted by $\mathbf{1}_n$. The vector $G\mathbf{1}_n$ represents the author connectivities, capturing each author's degree, which corresponds to the number of direct connections. Meanwhile, the entries of $G^k\mathbf{1}_n$ give the total count of k-length paths originating from each respective author.

The vector of Katz-Bonacich centralities is thus:

$$\mathbf{b}(g, \lambda) = \lambda G\mathbf{1}_n + \lambda^2 G^2\mathbf{1}_n + \lambda^3 G^3\mathbf{1}_n + \cdots = \sum_{k=0}^{\infty} \lambda^k G^k\mathbf{1}_n \cdot (\lambda G\mathbf{1}_n).$$

If λ is small enough, this infinite sum converges to a finite value, which is $(\mathbf{I} - \lambda G)^{-1}$, where $\mathbf{I}$ is the identity matrix of the proper size. The vector of Katz-Bonacich centralities can then be written as follows:

$$\mathbf{b}(g, \lambda) = (\mathbf{I} - \lambda G)^{-1} \cdot (\lambda G\mathbf{1}_n).$$

It's worth noting that, according to its definition, the Katz-Bonacich centrality of an author is zero in an empty network. Additionally, it becomes zero when $\lambda = 0$, and is increasing and convex in λ. Moreover, it is bounded from below by λ times the author's connectivity, expressed as $b_i(g, \lambda) \geq \lambda g_i$.
In the forthcoming section, I introduce a formal model that provides a clear micro-foundation for the Katz-Bonacich centrality. This model will guide us in formulating the following hypothesis:

$H3$: Authors' productivity is positively correlated with their Katz-Bonacich centrality.

A fundamental question in my analysis pertains to how productivity and the impact of knowledge spillovers on productivity are contingent upon the individual characteristics of scientists. There exists a well-established body of literature investigating the individual characteristics that influence scientists' productivity. When studying the interplay between these characteristics and the coauthorship network in which scientists operate, two natural hypotheses arise. First, I would expect to find the same qualitative results regarding the effect of scientists' characteristics on productivity. Second, in the absence of explicitly considering coauthors' influence in the analysis, there might be an overestimation of their effects. Certain characteristics (such as gender and experience) possess both direct and indirect effects on productivity, as they also serve as determinants of scientific connections. Consequently, if coauthorship connections are disregarded, estimates might incorporate indirect effects alongside direct ones. This leads us to the formulation of my next hypothesis:

$H4$: When incorporating coauthorship spillovers, the impact of scientists' characteristics on productivity will exhibit the same qualitative pattern as observed in analyses without such spillovers. Nevertheless, disregarding coauthorship spillovers would result in an overestimation of individual effects.

2.2 A Formal Framework

Consider an undirected network g where a finite set of authors $N = \{1, \ldots, n\}$ interact. G represents the $n \times n$ zero-diagonal symmetric adjacency matrix, with $g_{ij} = 1$ if author i has co-authored with author j, and $g_{ij} = 0$ otherwise. The row normalised adjacency matrix is denoted by $\bar{g}$. Given the network structure, authors decide on their production levels during a period. Let y_i denote the production level of author i, and $y = (y_1, \ldots, y_n)^\intercal$ represent the population's production profile in the network. Each author i chooses $y_i \geq 0$ and receives a utility $\mathcal{U}_i(\mathbf{y})$ given by:

$$\mathcal{U}_i(\mathbf{y}) = \left(\pi_i + \lambda \sum_{j=1}^{n} g_{ij} y_j \right) y_i - \frac{1}{2} y_i^2$$

where $\lambda \geq 0$. This utility function follows a standard cost-payoff structure akin to Lee et al. [2021], and it is additively separable into idiosyncratic production components and contributions from coauthors or peer effects. Each author incurs a cost associated with the process of producing science, represented by $\frac{1}{2} y_i^2$. The component π_i introduces exogenous heterogeneity, capturing observable differences between authors such as gender, past citations, experience in the field, but also the average characteristics of coauthors named *contextual effects*. Specifically,

$$\pi_i = \eta + X_i \beta + \bar{G}_i X \gamma + u_i$$

where X_i is a row vector of observable exogenous characteristics of author i, $\beta = (\beta_1, \ldots, \beta_k)'$ and $\gamma = (\gamma_1, \ldots, \gamma_k)'$ are vectors of coefficients, η is a constant term, and u_i represents the unobservable (to the econometrician) characteristics of author i.

The peer effect component also exhibits heterogeneity, with this endogenous diversity stemming from authors' varied positions within the coauthorship network g and the resulting production level. The second partial derivative of $\mathcal{U}_i(y, g)$ with respect to y_i and y_j, denoted as $\frac{\partial^2 \mathcal{U}_i(y,g)}{\partial y_i \partial y_j}$, is given by $\lambda g_{ij} \geq 0$. When authors i and j are coauthors, the positive value of the cross derivative λ signifies strategic complementarity in production. Conversely, if authors i and j are not coauthors, this cross derivative is zero. In the scientific context, a positive λ indicates that when authors i and j collaborate (i.e., $g_{ij} = 1$), an increase in j's production leads to a rise in i's marginal utility from production. I refer to λ as the co-author interaction coefficient.

2.3 Equilibrium and its Implications

Let's now describe the Nash equilibrium of the game, where each author $i = 1, \ldots, n$ selects simultaneously their own production level $y_i \geq 0$. In equilibrium, each author maximizes their utility, and their best-response function is given by:

$$y_i = \lambda \sum_{j=1}^{n} g_{ij} y_j + \eta + X_i \beta + \bar{G}_i X \gamma + u_i \tag{1}$$

Here, the author's outcome comprises two distinct effects: a network-specific effect and an idiosyncratic one. In other words, individual behavior can break down into two components: an exogenous part and an endogenous peer effect component that depends on the author under consideration.

Let $\mu_1(A)$ denote the largest eigenvalue of a square matrix A, and π denotes the non-negative n-dimensional vector, where each element corresponds to π_i. I denotes the $n \times n$ identity matrix, and $\mathbf{1}_n$ denotes the n-dimensional vector of ones. The following proposition define the Nash equilibrium in pure strategies.

Proposition 1 [Ballester et al., 2006]: If $|\lambda| \leq 1/\mu_1(G)$, then $I_n - \lambda G$ is nonsingular, and the network possesses a unique interior Nash equilibrium in pure strategies with the equilibrium production vector $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)^\intercal$ given by:

$$\mathbf{y}^* = \mathbf{y}^*(g) = b_\pi(g, \lambda), \quad \text{where} \quad b_\pi(g, \lambda)$$

is the weighted Katz-Bonacich centrality defined as:

$$b_\pi(g, \lambda) = (I_n - \lambda G)^{-1}\pi = M(g, \lambda)\pi = \sum_{k=0}^{\infty} \lambda^k G^k \pi.$$

When all the agents are isolated, meaning the network is empty, we have $b_\pi(\emptyset, \lambda)_i$ for all i, and thus $y_i^*(y_{-i}, \emptyset) = \alpha + X_i\beta$. In this scenario, there are no multiplier peer-effects, and production solely depends on each author i's idiosyncratic characteristics.

The condition $\mu_1(G)\lambda < 1$ in Proposition 1 implies that the impact of network complementarities must be sufficiently small relative to the individual convexity of costs. This limitation prevents positive feedback loops triggered by such complementarities from escalating without bound. It's essential to note that this condition doesn't directly constrain the absolute values of these cross-effects, but rather their relative magnitude. Network complementarities are assessed through the compound index $\lambda\mu_1(G)$, where λ represents the intensity of each non-zero cross-effect, while $\mu_1(G)$ captures the overall pattern of these positive cross-effects.

Proposition 1 establishes a connection between an author's equilibrium production and her network position, as quantified by her Katz-Bonacich centrality.

Model discussion In this model, the structure of the social network, particularly authors' positions within it, serves as the primary explanatory variable for their production, alongside idiosyncratic heterogeneity. This linkage, termed the Nash-Katz-Bonacich linkage, highlights the significance of network structure in shaping individual outcomes. While previous research in the economics of science has underscored the importance of individual characteristics and peer effects, it often overlooks the macro structure of the network.

The novelty of this model lies in its emphasis on network structural properties as fundamental to understanding peer influence on individual production. Specifically, within the realm of linear-quadratic utility functions, the Katz-Bonacich centrality index effectively captures peer effects within networks. In the subsequent sections, I test the empirical relevance of this model.

The derived empirical measure of peer effects diverges notably from prior studies in this domain. Rather

than examining group peer effects on an individual's scientific productivity, I focus on discerning the impact of network effects distinct from idiosyncratic characteristics. This allows to gauge the significance of network effects on scientific output, particularly to elucidate how an author's network position influences her output, once individual characteristics are considered.

3 Econometric model

Assume that we observe data from $\bar{c}$ communities $c = \{1, \ldots, \bar{c}\}$ each comprised of n_c authors and characterized by a network $G_c = \{g_{ij,c}\}$. Moreover, assume the vector of characteristics of authors in community c, $\boldsymbol{\pi}_c = (\pi_{1,c}, \ldots, \pi_{n,c})^\intercal$, is a linear function of a vector of the authors' characteristics in community c:

$$\boldsymbol{\pi}_c = \eta_c 1 + X_c\beta + \bar{\mathbf{G}}_c X_c \gamma + \mathbf{u}_c$$

The specification of the econometric model follows the equilibrium best-reply function of the network game (1) so that it has a clear microfoundation.

$$y_{i,c} = \lambda \sum_{j=1}^{n_c} g_{ij,c} y_{j,c} + x_{i,c}^{'}\beta + \sum_{j=1}^{n_c} \bar{g}_{ij,c} x_{j,c}^{'}\gamma + \eta_c + u_{i,c},$$

for $i = 1, \cdots, n_c$ and $c = 1, \cdots, \bar{c}$. Let $\mathbf{Y}_c = (y_{1,c}, \cdots, y_{n,c})^{'}$, $\mathbf{X}_c = (x_{1,c}, \cdots, x_{n,c})^{'}$, and $\mathbf{u}_c = (u_{1,c}, \cdots, u_{n,c})^{'}$. Then, (3) can be written in matrix form as

$$\mathbf{Y}_c = \lambda \mathbf{G}_c Y_c + \mathbf{X}_c\beta + \bar{\mathbf{G}}_c \mathbf{X}_c \gamma + \eta_c \iota_{n,c} + \mathbf{u}_c$$

Let $\text{diag}(A_{ij})$ denote a block diagonal matrix in which the diagonal blocks are $m_j \times n_j$ matrices $A_j s$. For a data set with $\bar{c}$ communities, let $\mathbf{Y} = \left(\mathbf{Y}_1^{'}, \cdots, \mathbf{Y}_{\bar{c}}^{'}\right)^{'}$, $\mathbf{X} = \left(\mathbf{X}_1^{'}, \cdots, \mathbf{X}_{\bar{c}}^{'}\right)^{'}$, $\mathbf{u} = \left(\mathbf{u}_1^{'}, \cdots, \mathbf{u}_{\bar{c}}^{'}\right)^{'}$, $\eta = \left(\eta_1^{'}, \cdots, \eta_{\bar{c}}^{'}\right)^{'}$, $\mathbf{G} = \{G_c\}_{c=1}^{c=\bar{c}}$, $\bar{\mathbf{G}} = \{\bar{\mathbf{G}}_c\}_{c=1}^{c=\bar{c}}$ and $\mathbf{L} = \{\iota_{n,c}\}_{c=1}^{c=\bar{c}}$. The general econometric model can be written as

$$\mathbf{Y} = \lambda \mathbf{G} Y + \mathbf{X}\beta + \bar{\mathbf{G}}\mathbf{X}\gamma + \mathbf{L}\eta + \mathbf{u}$$

The network-specific parameters η are allowed to depend on $\mathbf{G}$, $\bar{\mathbf{G}}$ and $\mathbf{X}$ as in a fixed effect panel data model. To avoid the incidental parameter problem when the number of groups is large, I eliminate the term $\mathbf{L}\eta$ using the deviation from group mean projector, where $\mathbf{J}_c = \mathbf{I}_{n_c} - \frac{1}{n_c}\iota_{n_c}\iota_{n_c}^{'}$. This transformation is analogous to the within transformation for a fixed effect panel data model. As $\mathbf{JL} = 0$, the transformed network model is

$$\mathbf{JY} = \lambda \mathbf{JG} Y + \mathbf{JX}\beta + \mathbf{J}\bar{\mathbf{G}}\mathbf{X}\gamma + \mathbf{JL}\eta + \mathbf{Ju} \tag{2}$$

The error term u is assumed to be independently distributed but allowed to be heteroscedastic, where $\boldsymbol{\Sigma} = E(uu^\intercal) = \text{diag}\{\sigma_i^2\}$ represents a diagonal matrix. The econometric model (2) corresponds to a higher-order spatial autoregressive (SAR) model [Lee et al., 2021]. Assessing the effects of co-author production on individual output, particularly identifying the endogenous co-author effect (λ), is often fraught with

econometric challenges. These issues, well-documented in the literature, include correlated or common shock effects and the endogenous sorting of individuals into groups [Manski, 1993].

3.1 Threats to Identification

Correlated or common-shock effects Correlated or common-shock problems arise in network models because authors within the same network may behave similarly due to shared environmental factors. These factors can potentially confound the co-author effect λ I aim to identify. One approach to address this issue is by introducing community fixed effects.

Introducing community fixed effects raises an interpretation issue: If we assume that differences in publication records between communities across the entire network primarily reflect overall trends and the scale of the field, rather than genuine disparities in academic productivity and talent, then these differences should be controlled for with network fixed effects. This approach focuses on the variability in network position independently of community distinctions. Conversely, if one believes that a higher volume of publications in a field across the entire network genuinely indicates higher productivity, then field fixed effects should not be included in the model specification.

I align with the former viewpoint and include community fixed effects. This perspective is commonly adopted in empirical analyses of peer effects. Existing literature consistently incorporates network fixed effects, which help estimate peer effects after accounting for the structural composition of the community. While network fixed effects allow differentiation between endogenous and correlated effects, they may not fully capture the causal influence of peers on individual productivity.

Endogeneity of the coauthorship network The main econometric challenge arises from the endogenous formation of coauthorship relationships. Authors make choices regarding whom to collaborate with, and these decisions may be influenced by unobservable factors. For instance, an author might opt for collaboration to tackle complex ideas collectively or because they prefer working with individuals who share similar characteristics or scientific abilities. In particular, a high assortativity in the matching process is observed in the scientific network, indicating that less productive authors tend to collaborate primarily with others of similar abilities. If this selection was disregarded, its impact would be inaccurately allocated to collaboration, leading to biased coefficient estimates.

3.2 Estimation procedure

When the adjacency matrix $\mathbf{G}$ is considered exogenous, a 2SLS estimator can be use for the estimation of the linear social interaction models described by Equation (2). More specifically, the 2SLS relies on the linear moment condition $\mathbf{Z}'\mathbf{J}\mathbf{u}(\theta) = 0$, where $\mathbf{Z}$ is a matrix of instrumental variables (IVs) composed of linearly independent columns, namely $\mathbf{J}[\mathbf{X}, \bar{\mathbf{G}}\mathbf{X}, \mathbf{G}\mathbf{X}, \bar{\mathbf{G}}^2\mathbf{X}]$, let also $\underline{\mathbf{X}} = \mathbf{J}[\mathbf{G}\mathbf{Y}, \mathbf{X}, \bar{\mathbf{G}}\mathbf{X}]$ and

$$\mathbf{J}\mathbf{u}(\theta) = \mathbf{J}\mathbf{Y} - \underline{\mathbf{X}}\theta$$

To tackle this issue of network endogeneity, I used a 2SLS estimator with a predicted adjacency matrix to estimate Equation (2). This approach entails substituting the observed adjacency matrix $\mathbf{G}$ with a predicted counterpart $\hat{\mathbf{G}}$ derived from exogenous covariates [Kelejian and Piras, 2014].

As with any instrumental variables approach, the effectiveness of this method in addressing endogeneity hinges on the availability of a reliable instrument — a variable that strongly predicts the network covariate while remaining orthogonal to the outcome equation. This task poses specific challenges in this context, where scientific collaboration is largely driven by the scientists' objectives. Nevertheless, identifying such an instrument becomes more feasible under external constraints that limit link formation despite potential benefits.

In this study, I used the overlap in research interest in the period before the collaboration is taking place. This overlap serves as a plausible instrument under the following assumptions: it is exogenous to the scientific knowledge production process and remains relevant, even if a researcher initiates a new line of research.

Next, I outline the procedure for obtaining a predicted coauthorship matrix in the first stage. I begin by estimating a logistic regression model to model the relationship between the elements of the adjacency matrix g_{ij} and the dyadic covariates W_{ij}. The logistic function is employed to model this relationship:

$$\hat{g_{ij}} = \frac{\exp(\hat{\delta_0} + W_{ij}\hat{\delta_1})}{1 + \exp(\hat{\delta_0} + W_{ij}\hat{\delta_1})} \tag{3}$$

Here, $\hat{\delta_0}$ and $\hat{\delta_1}$ are obtained from a logistic regression of g_{ij} on W_{ij}. To ensure that the predicted adjacency matrix is uniformly bounded in row and column sums, I normalize $\hat{g_{ij}}$ by dividing it by $\hat{d} = \max\{\max_i \sum_{j=1}^{n} \hat{g_{ij}}, \max_j \sum_{i=1}^{n} \hat{g_{ij}}\}$ [Kelejian and Prucha, 2010]. Subsequently, I define the element $(\hat{G})_{ij}$ of the predicted adjacency matrix $\hat{G}$ as $\hat{g_{ij}}/\hat{d}$ if $i \neq j$ and zero otherwise. This normalization procedure ensures that the predicted adjacency matrix is appropriately scaled and maintains necessary properties for the subsequent analysis.

For the average characteristics of peers ($\bar{\mathbf{G}}\mathbf{X}$) of the econometric model (2), I rely on the IV strategy proposed by Jochmans [2023]. The latter instrument is constructed as follows: for each scientist, I create a subnetwork by removing all links involving that scientist. This "leave-own-out" network is exogenous and provides valuable predictive information about the individual's linking behavior if certain conditions are met. Specifically, the self-selection problem must be due to agent-specific unobservables that do not influence the decision of any other pair of agents to form a link. These unobservables can be dependent within groups and may even cause group-level endogeneity. The conditions also accommodate interdependent link formation, as long as this interdependency is confined to within-group interactions. I then instrument the average coauthor characteristics using the average characteristics within the leave-own-out network.

The construction of instrumental variables for $\bar{\mathbf{G}}$ with weights coming from $\mathbf{G}_{-i}$.

$$(H_{-i})_{i',j} = \begin{cases} \frac{(G)_{i',j}}{\sum_{j' \neq i}(G)_{i',j'}} & \text{if } i' \neq i \text{ and } j \neq i \text{ and } \sum_{j' \neq i}(G)_{i',j'} > 0 \\ 0 & \text{otherwise} \end{cases}$$

This is the row-normalized version of the previously introduced adjacency matrix $\mathbf{G}_{-i}$. It has been supplemented with an additional row and an additional column of zeros, each inserted at position i. This augmentation is performed for notational convenience, ensuring that the resulting matrices retain the $n \times n$ dimensions, matching the size of the matrix H for the complete network. The matrix H_{-i} represents the transition matrix for the network with all connections involving agent i removed.

For each scientist we set up the subnetwork obtained on removing all link decisions in which this scientist is involved. This leave-own-out network is exogenous and contains useful predictive information about the scientist's own link choices. One can interpret $\bar{\mathbf{G}}$ as a transition matrix giving the probability of ending at scientist j, from scientist i, in a single step in the network defined by the original adjacency matrix $\mathbf{G}$. The entries of the $n \times n$ matrix

$$(Q_1)_{i,j} = \frac{1}{n-1} \sum_{i' \neq i} (H_{-i})_{i',j} \,,$$

in contrast, give the probability of arriving at scientist j in the network defined by $\mathbf{G}_{-i}$, no matter the starting point, in a single step. In full analogy to Q_1, the entries of the $n \times n$ matrix

$$(Q_2)_{i,j} = \frac{1}{n-1} \sum_{i' \neq i} \sum_{j'=1}^{n} (H_{-i})_{i',j'} \, (H_{-i})_{j',j} \,,$$

give the probability of arriving at scientist j in the network defined by $\mathbf{G}_{-i}$, no matter the starting point, in two steps. $\bar{\mathbf{G}}\mathbf{X}$ can be instrumented by $\mathbf{Q}_1\mathbf{X}$ and $\mathbf{Q}_2\mathbf{X}$. The IV matrix based on the predicted adjacency matrix $\hat{\mathbf{G}}$, $\mathbf{Q}_1$, $\mathbf{Q}_2$, is denoted by $\hat{\mathbf{Z}}$ and includes linearly independent columns of $J[X, \mathbf{Q}_1\mathbf{X}, \mathbf{Q}_2\mathbf{X}, \hat{\mathbf{G}}\mathbf{L}]$. The corresponding 2SLS estimator is given by

$$\hat{\theta}_{2sls} = [\underline{\mathbf{X}}' \hat{\mathbf{Z}} (\hat{\mathbf{Z}}' \hat{\mathbf{Z}})^{-1} \hat{\mathbf{Z}}' \underline{\mathbf{X}}]^{-1} \underline{\mathbf{X}}' \hat{\mathbf{Z}} (\hat{\mathbf{Z}}' \hat{\mathbf{Z}})^{-1} \hat{\mathbf{Z}}' \mathbf{J}\mathbf{Y} \tag{4}$$

As pointed out by Lee et al. [2021], the consistency of the proposed GMM estimator does not rely on the consistency of the estimator $\hat{\delta}$. Suppose $\hat{\delta}$ converges in probability to a well defined limit δ^* such that $\sqrt{n}(\hat{\delta} - \delta^*) = \mathcal{O}_p(1)$. Let $\mathbf{G}^*$ be defined in the same fashion as $\hat{\mathbf{G}}$, with $\hat{g_{ij}}$ in $\hat{\mathbf{G}}$ replaced by

$$g^*_{ij} = \frac{\exp(\delta^*_0 + W_{ij}\delta^*_1)}{1 + \exp(\delta^*_0 + W_{ij}\delta^*_1)}.$$

Then, the parameters in Equation (3) can be identified via the IV matrix $\mathbf{Z}^*$ consisting of linearly independent columns of $\mathbf{J}[\mathbf{X}, \mathbf{Q}_1\mathbf{X}, \mathbf{Q}_2\mathbf{X}, \mathbf{G}^*\mathbf{L}]$. Under some regularity conditions, it is possible to show the asymptotic equivalence between $\hat{\mathbf{Z}}' \mathbf{J}\mathbf{u}(\theta) = 0$ and the infeasible linear moment condition $\mathbf{Z}^{*'} \mathbf{J}\mathbf{u}(\theta) = 0$, by consequence $\hat{\theta}_{2sls}$ defined in Equation (4) is root-n consistent and asymptotically normal.[1]

The instrumental variables (IVs) derived from the predicted adjacency matrices may be weak if the dyadic characteristics W_{ij} do not effectively explain the formation of network links. Lee [2007] suggests extending the 2SLS method to a comprehensive GMM framework with additional quadratic moment conditions based on the covariance structure of the reduced form equation. This approach can enhance both identification and

[1] The asymptotic distribution of $\hat{\theta}_{2sls}$ and $\hat{\theta}_{gmm}$ are given in Appendix.

estimation efficiency. To address the potential weak IV problem, I introduce a quadratic moment condition for the estimation of Equation (2).

Given that $\mathbf{G}^*$ is exogenous and has a zero diagonal, it follows that $\mathbb{E}((\mathbf{Ju})^{'}\mathbf{G}^*\mathbf{Ju}) = \text{tr}(\mathbf{G}^*) = \text{tr}(\mathbf{G}^*\text{diag}(\sigma_i^2)) = 0$. This implies an infeasible quadratic moment condition $(\mathbf{Ju}(\theta))^{'}\mathbf{G}^*\mathbf{Ju}(\theta) = 0$. The feasible quadratic moment condition $(\mathbf{Ju}(\theta))^{'}\hat{\mathbf{G}}\mathbf{Ju}(\theta) = 0$ is asymptotically equivalent to $\mathbf{Ju}(\theta)'\mathbf{G}^*\mathbf{Ju}(\theta) = 0$. Combining the linear and quadratic moment conditions, the GMM estimator is defined as follows:

$$\hat{\theta}_{gmm} = \arg\min h(\theta)^{\intercal}\hat{\Omega}^{-1}h(\theta), \tag{5}$$

where $h(\theta) = \left[\mathbf{Ju}(\theta)^{\intercal}\hat{\mathbf{Z}}, \mathbf{Ju}(\theta)^{\intercal}\hat{G}\mathbf{Ju}(\theta)\right]^{\intercal}$ and $n^{-1}\hat{\Omega}$ is a consistent estimator of the variance-covariance matrix of the moment function $h(\theta)$.

4 Network and Data

This section is divided in four parts. In the first part, I describe the datasets used to construct my database. In the second part, the construction of the sample is presented with an emphasis on network definition. In the third part, I document the construction of the variables. Ultimately, I provide descriptive statistics.

4.1 Data Sources

Publication Records The primary data source for this analysis is the MEDLINE publication database.[2] MEDLINE, maintained by the US National Library of Medicine, is the leading bibliographic database for life sciences, with nearly all journal articles dating back to 1946. It includes detailed information on articles such as author names, publication dates, journal titles, grant acknowledgments, and Medical Subject Headings (MeSH) terms. MeSH terms are a controlled vocabulary that offers a highly granular classification of biomedical research, assigned by professional indexers focused solely on the scientific content.[3] This ensures a more objective description of each paper. Citation data are obtained from the NIH Open Citation Collection [Hutchins et al., 2019]. A detailed procedure for identifying relevant cancer-related publications is provided in the appendix.

Author Name Disambiguation Author names are not reliable unique identifiers, as different individuals may share the same name, and names or affiliations can change over time. This creates a challenge for author-specific publication data collection, especially since bibliographical databases typically lack individual author identifiers. In Medline, nearly two-thirds of authors have ambiguous names, with the same last name and first initial shared by multiple authors [Smalheiser et al., 2009]. The "John Smith" issue has prompted the development of various disambiguation algorithms aimed at identifying individuals based on

[2] I focus on research papers, excluding book reviews, letters, and similar non-research content.

[3] MeSH terms are assigned by specialists who are experts in specific health science areas and work in coordination with various specialized vocabularies.

author names.[4]

Despite efforts to create global author IDs (e.g., ORCID, ResearcherID), many Medline articles, especially those published before 2003 (before the ORCID field was added to PubMed), offer limited author information—just the last name, first initial, and, for first authors prior to 2014, affiliation.

Xu et al. [2020] developed a PubMed Knowledge Graph (PKG) that improves author name disambiguation. Their algorithm, which clusters articles to infer authorship, has proven highly accurate, particularly for NIH-funded scientists. This approach builds on the probabilistic disambiguation model by Torvik et al. [2005], which assumes that articles by the same author share more common attributes than those by different authors.

Author Attributes To complement the coauthorship and productivity data, I use metadata from the Torvik research group to create author-level variables such as ethnicity, gender, and affiliation.

- **Ethnicity**: Determined using the Ethnea tool, which assigns ethnicity based on authors' first and last names. Unlike other algorithms, Ethnea prioritizes nationality over distant ancestry and accounts for dual ethnicities, which are common due to marriage, migration, and assimilation.[5]
- **Gender**: Assigned using the Genni tool, which probabilistically matches a gender to a first name, considering the ethnicity linked to the last name. For instance, the name "Andrea" is associated with a female gender if the last name is "FRENCH" but with a male gender if it is "ITALIAN."
- **Affiliations**: Derived from the MapAffil tool, which links PubMed author affiliation strings to cities and geocodes worldwide. The dataset, based on a snapshot of PubMed in October 2016, includes affiliations for articles prior to 2014 (when only first author affiliations were recorded). It also incorporates records from PubMed's supplement sources, such as PMC, NIH grants, the Microsoft Academic Graph, and the Astrophysics Data System. One limitation is that it does not cover data after 2015, capturing 62.9% of affiliation instances in PubMed.

For each author, I also record their degree (including type and date) and whether they are a prize-winning scientist.

4.2 Network Construction

My analysis focuses on authors in cancer research, where collaboration is common and productivity can be measured by publications. Scientific collaborations are crucial for knowledge exchange and the generation of new ideas, as much new knowledge remains tacit and confined within tightly-knit groups. To ensure the

[4]Author name disambiguation remains a key unresolved issue in bibliometrics. Kang et al. [2009] provides an excellent review of the literature and approaches for addressing it.

[5]The Ethnea tool from the University of Illinois determines authors' ethnicity based on their names. It is available at http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py.

Table 1: Summary Statistics

Variables	N	Mean	St. Dev.	Min	Max
Asinh production 2010-2014	1,455	2.8	1.4	0.1	7.1
Asinh Citations pre-period	1,455	5.6	3.4	0.0	11.9
Male	1,455	0.7	0.4	0	1
Early Career Prize	1,455	0.01	0.1	0	1
Clinical concentration	1,455	0.1	0.2	0.0	1.0
Phd	1,455	0.4	0.5	0	1
Md	1,455	0.7	0.5	0	1
Decades after graduation	1,455	1.8	1.1	-1.2	5.8
Late Career Recognition	1,455	0.05	0.2	0	1
Ivy League Undergrade	1,455	0.2	0.4	0	1

sample includes only active researchers, I restricted it to scientists who graduated before 2014 (using the earliest graduation date for those with dual degrees) and were no longer undergraduate students during the study period. Additionally, I required each author to have at least one publication as a first or last author in the field. This aligns with a common social norm in life sciences, where the last author is typically the principal investigator, the first author is the junior researcher responsible for the work, and other authors are credited in decreasing order [Zuckerman, 1968, Nagaoka and Owan, 2014].

In the coauthorship network, two authors are linked if they co-authored at least one paper in cancer research from 2010 to 2014. This network includes a giant component and many smaller components.

To define tightly-knit groups within the giant component, I used a modularity-based algorithm [Blondel et al., 2008], which partitions the network to minimize the modularity criterion.[6] To avoid a many-IV issue, I included only communities with at least 30 authors. These communities were then pooled into a single network, with each community forming a component of the overall structure.

4.3 Variable Definitions

IV Construction To measure homophily among coauthors, I build on Boschma [2005], assessing proximity across cognitive, social, institutional, and cultural dimensions, excluding geographical proximity due to its correlation with institutional proximity.

- **Cognitive Proximity**: Reflects shared knowledge and research interests, influencing collaboration

[6] Modularity-based algorithms are well discussed in the computer science and physics literature (e.g., Loscalzo and Barabási [2016], Ch. 9).

likelihood. I measure research interests using MeSH terms from cancer articles published in the last five years and compute similarity via cosine similarity [Breschi et al., 2003].

- **Social Proximity**: Based on past collaborations, considering both direct and indirect connections over the last five years [Uzzi, 1996, Breschi and Lissoni, 2009, Fafchamps et al., 2010]. I focus on whether coauthors i and j are directly connected or share a mutual co-author.
- **Institutional Proximity**: Highlights collaborations within the same institution, reflecting the role of face-to-face interactions [Dahlander and McFarland, 2013, Long et al., 2014]. I use MapAffil tools to identify main institutional affiliations.
- **Cultural Proximity**: Captures shared language or social norms, measured by authors' ethnicity using the Ethnea tool, aligning with studies showing higher collaboration rates among ethnically similar individuals [Dahlander and McFarland, 2013, Freeman and Huang, 2015].
- **Gender**: I also consider gender, which can influence collaboration due to homophily, discrimination, or gender-specific risk aversion [Ductor et al., 2023].

Author Productivity I gather demographic data for each scientist from their CVs, supplemented by professional social networks and faculty web pages. This includes degrees (bachelor's, MD, PhD, MD/PhD) and institutions where they were earned, with gender determined using Torvik Research Group metadata for instrumental variable construction.

I compute the **cumulative number of citations**, weighted by impact factor, for each scientist up to the start of the period. To identify **early-career prize winners**, I include recipients of prestigious scholarships (Pew, Searle, Beckman, Rita Allen, Packard) awarded from 1981 to 2014. These scholarships are granted to 20–40 young life scientists annually, providing early recognition in their independent careers.

For **late-career recognition**, I include scientists elected to the National Academy of Sciences and the Institute of Medicine from 1970 to 2014, as well as current and former Howard Hughes Medical Investigators (HHMIs). HHMI selects a distinguished group of mid-career biomedical scientists every three years, recognizing their potential for significant contributions.

5 Empirical findings

In this section, I first discuss the empirical relevance of the instrument used in my 2SLS estimation procedure. This is followed by the presentation of the estimation results, after which I provide robustness checks.

5.1 Relevance of the Instrument

I begin my empirical analysis by discussing the estimation results of the dyadic network formation model used to instrument the adjacency matrix. This exercise constitutes the formal first step of the approach de-

scribed earlier. Table (2) presents the estimation results of Model (3), in which I include all the explanatory variables previously highlighted by the underlying literature (see discussion in the previous section).

I find that attending the same graduate institution during the same period has no significant effect on the likelihood of collaboration. This may reflect an intentional effort (from institutions) to guide students into distinct research lines, possibly to reduce competition among graduates from the same institution on the job market. The remaining variables show expected relationships.

With regard to the social dimension, I found that being closer in the previous coauthorship network (located at distance 1 or 2) facilitates the formation of a collaboration, so sharing a certain proximity in the network of past coauthors significantly increases the intensity of a collaboration. This result confirms the relationship between social proximity and collaboration intensity found in previous studies. One possible explanation for this result is that the network is a source of information about the quality of potential coauthors, particularly with respect to their respective productivity.

The material resources needed to conduct research in the medical field provide us with a key to interpreting institutional homophily. As this type of research is essentially experimental, researchers are very likely to be dependent on expensive infrastructures such as hospitals, genome sequencing equipment, etc. The fact that both researchers are located within the same institution makes it much easier for them to conduct their research together, which is a plausible explanation for the higher intensity of the collaboration. A second type of argument can also be put forward: the co-location of two researchers within the same institution facilitates informal exchanges that may lead to the formalisation of a collaboration by co-authoring an article. It would be interesting to disentangle the impact of each of these channels.

With respect to cultural proximity, I found that sharing the same ethnicity is a significant predictor of collaboration intensity. This finding is consistent with previous results documenting the homophily of individuals on this dimension.

Finally, one explanation for homophily on the seniority, known as (positive) assortativity, is that nodes are born at different times, introducing certain correlations as a function of age. For instance, academics initiate their research careers at different points in time. Older researchers have had more opportunities to collaborate with other researchers (tending to give them a higher degree) and also relatively more opportunities to collaborate with other older researchers, producing a positive correlation in the seniority of coauthors.

5.2 In the quest of peer effects

As a benchmark, Column 1 in Table 7 reports the OLS estimates of the model without network effect, in which scientific productivity is explained using only scientists' characteristics (ignoring the fact that scientists are connected). While the coefficient for the male variable is positive across all models, it is not statistically significant, implying no clear gender effect on citation counts. Holding a PhD, on the other hand, shows a robust and significant positive effect in all models, indicating that PhD holders tend to garner more citations compared to their peers without a PhD. Interestingly, holding an MD does not appear to have a significant

Table 2: Logistic regression of link formation.

Cosine Similarity	4.919^{***}
	(0.145)
Past coauthor	2.676^{***}
	(0.042)
Past common coauthor	1.134^{***}
	(0.038)
Same gender	0.125^{***}
	(0.031)
Same Ethnicity	0.068^{**}
	(0.031)
Experience Difference	-0.059^{***}
	(0.015)
Same Grad school	0.087
	(0.181)
Same affiliation	1.230^{***}
	(0.031)
Observations	132,916
Log Likelihood	−17,466.310
Akaike Inf. Crit.	34,950.620
McFadden Pseudo R^2	0.296

Note Results for Model (3) of the article are displayed. The dependent variable is defined as the existence of collaboration betwen author i and author j. The independent variables capture differences in characteristics between i and j. A precise definition of the variables at the individual level can be found in the previous section. Standard errors are reported in parentheses. An intercept is included. $*$, $**$, and $***$ indicate statistical significance at the 10%, 5%, and 1% levels.

impact, with coefficients close to zero in most models. Indeed except in some cases most of the MDs are not fully dedicated to research since they also hold some clinical position which makes them less endowed with time for research. Moreover, winning an early career prize is negatively associated with productivity, though the results are non significant.

Columns 2-7 in Table 7 present the estimate of model (2) using both 2SLS and GMM controlling for network endogeneity (2SLS-2 and GMM). The estimates reveal a positive and statistically significant estimate of λ, which suggests the presence of network externalities in line with Hypothesis 1. The coefficients on most of the control variables retain the same sign and statistical significance across models, columns 1–7. The interpretation of the coefficients and their magnitude differs depending on whether or not the model takes into account the network. Indeed, if $\lambda > 0$, then the marginal effect of the $k - th$ covariate is not β_k but

Table 3: Results for the period 2010-2014

	OLS	2SLS-1	2SLS-2	GMM	2SLS-1	2SLS-2	GMM
λ		0.034***	0.043***	0.044***	0.034***	0.044***	0.043***
		(0.002105)	(0.003254)	(0.010546)	(0.002142)	(0.003343)	(0.010802)
Asinh Citations pre-period	0.1392***	0.11***	0.101***	0.1***	0.11***	0.102***	0.102***
	(0.0141)	(0.013699)	(0.014)	(0.008447)	(0.01364)	(0.013898)	(0.008244)
Male	0.1055	0.076	0.065	0.105*	0.078	0.07	0.099*
	(0.0727)	(0.067781)	(0.068691)	(0.044883)	(0.067713)	(0.068103)	(0.040683)
Early Career Prize	-0.2564	-0.322	-0.362	-0.338	-0.275	-0.281	-0.231
	(0.3176)	(0.277158)	(0.290506)	(0.189667)	(0.279659)	(0.294893)	(0.301875)
Clinical concentration	0.8323***	0.312	0.277	0.483***	0.396*	0.266	0.344**
	(0.2289)	(0.203341)	(0.208594)	(0.116477)	(0.199537)	(0.199593)	(0.113489)
Phd	0.4661***	0.421***	0.405***	0.433***	0.445***	0.439***	0.455***
	(0.0941)	(0.087583)	(0.091484)	(0.061812)	(0.086537)	(0.087309)	(0.0632)
Md	0.279***	0.112	0.06	0.02	0.103	0.05	0.023
	(0.1014)	(0.096707)	(0.102343)	(0.070568)	(0.096481)	(0.098611)	(0.06726)
Decades after graduation	-0.243***	-0.202***	-0.188***	-0.197***	-0.205***	-0.193***	-0.202***
	(0.0451)	(0.042798)	(0.043236)	(0.026871)	(0.043049)	(0.043475)	(0.027784)
Late Career Recognition	0.9209***	0.663***	0.552**	0.783***	0.674***	0.6***	0.687***
	(0.1599)	(0.162825)	(0.176417)	(0.155259)	(0.163082)	(0.172007)	(0.167128)
Ivy League Undergrade	0.2344***	0.163*	0.152	0.196***	0.165*	0.144	0.164**
	(0.0879)	(0.079496)	(0.081415)	(0.058465)	(0.080044)	(0.081405)	(0.062943)
Cragg-Donald Wald F statistic	-	14.9358	10.2753	-	28.9281	14.511	-
OIR test p-value	-	0.2402	0.9416	1	0.9818	0.9836	1
Community fixed-effect	yes	yes	yes	yes	yes	yes	yes
Contextual variables	no	yes	yes	yes	no	no	no

Notes: Estimates are derived from three models: 2SLS-1, which treats the adjacency matrix **G** as exogenous; 2SLS-2, which uses the predicted adjacency matrix as an instrument for the observed adjacency matrix **G**; and GMM, which adds a quadratic moment to the instrument used in 2SLS-2. The dependent variable is the number of citations over the study period, weighted by the journal impact factor. The following dummy variables are included: Male, Early Career Prize, PhD, MD, Late Career Recognition, and Ivy League Undergrade. The Clinical Concentration variable represents the proportion of clinical papers published by the author during the period. Experience is captured by the variable Decades after Graduation, based on the author's last degree (MD or PhD). Past Production is measured by the inverse hyperbolic sine (asinh) of previous citations, weighted by the impact factor of citing journals. Heteroscedasticity-robust standard errors are reported in parentheses. Statistical significance is indicated by $*$, $**$, and $***$ indicate statistical significance at the 10%, 5%, and 1% levels.

$(\mathbf{I} - \lambda\mathbf{G})^{-1}(\mathbf{I}\beta_k)$, which is an $n_r \times n_r$ matrix with its $(i,j) - th$ element representing the effect of a change in x_{jk} on y_i. The important difference is that the marginal effects are heterogeneous across individuals, since they depend on the individual's position in the network. This suggests that traditional estimates of the effects of these characteristics on author output which ignore these externalities may overestimate their importance by omitting a relevant variable, namely connectedness within the co-author network.

Variable	**No Network Effects**	**With Network Effects**					
	Panel (a)	**Panel (b)**					
		Direct Effects			**Indirect Effects**		
		Mean	**Min**	**Min**	**Max**	**Min**	**Max**
Asinh Citations pre-period	0.1392	0.1018	0.1007	0.1137	0.0000	0.0000	0.0148
Male	0.1055	0.0700	0.0693	0.0782	0.0000	0.0000	0.0102
Early Career Prize	-0.2564	-0.2852	-0.3186	-0.2823	-0.0001	-0.0414	0.0000
Clinical concentration	0.8323	0.2515	0.2490	0.2809	0.0001	0.0000	0.0365
Phd	0.4661	0.4432	0.4386	0.4950	0.0001	0.0000	0.0643
Md	0.2790	0.0438	0.0434	0.0489	0.0000	0.0000	0.0064
Decades after graduation	-0.2430	-0.1939	-0.2166	-0.1920	-0.0000	-0.0281	0.0000
Late Career Recognition	0.9209	0.5976	0.5915	0.6675	0.0002	0.0000	0.0867
Ivy League Undergrade	0.2344	0.1430	0.1416	0.1598	0.0000	0.0000	0.0207

Table 4: Notes. Panel (a) reports the OLS estimates. Panel (b) reports the effects with network considerations (Model 6 in table (7)).

In Table 4, I show the magnitudes of the diagonal elements of this matrix and compare them with the OLS marginal effects. As conjectured in Hypothesis 4, the evidence presented in this table suggests that usual estimates of the effects of these characteristics on scientific productivity that ignore those externalities risk overstating their importance by omitting a relevant variable, connectedness in the coauthorship network.
Late-career recognition significantly boosts citation counts, suggesting that academic honors in the later stages of a researcher's career enhance visibility and impact. Similarly, past productivity has a positive and significant effect on current productivity, indicating that researchers with a strong citation history tend to accumulate more citations over time. These findings suggest that both late-career recognition and past productivity contribute substantially to citation counts, with the former enhancing visibility and influence, and the latter reflecting a cumulative advantage in academic visibility.
However, these findings raise concerns about the accuracy of citation counts as the primary outcome variable. Although citations are adjusted for quality by weighting them with the journal impact factor (JIF), this adjustment also reflects the journal's prestige and the author's standing, rather than purely the scientific merit of the work. This introduces potential status effects, where an author's position in the academic hierarchy influences recognition, regardless of research quality. This dynamic complicates the distinction between genuine scientific contribution and status-driven recognition, as status can bias evaluations of research quality. Specifically, if the peer effect parameter λ is driven by the "Matthew effect" [Azoulay et al., 2014], where prior recognition leads to further success, it could undermine the validity of the theory explaining

peer effects, making policy implications based on it potentially flawed.

To account for these status effects, I use a method developed by Balzer and Benlahlou [2024], which predicts citations based on the textual characteristics of articles, independent of journal prestige or author status. This approach provides a more accurate estimate of an article's citation potential based solely on its scientific content. Even with this conservative measure of scientific value, the results (columns 1–6 of Table (5)) show a positive and significant estimate of λ in all specifications except for the GMM, providing evidence of genuine scientific peer effects. While conservative, this method represents a significant step toward isolating scientific merit from prestige in peer effect studies, without relying on natural experiments.

Table 5: Robustness check 2010-2014

	Predicted Weighted number of citations							Weighted number of citations
	OLS	2SLS-1	2SLS-2	GMM	2SLS-1	2SLS-2	GMM	2SLS-2
λ		0.013***	0.02***	0.037	0.013***	0.02***	0.02	0.0417***
		(0.001015)	(0.001601)	(0.029214)	(0.001027)	(0.001648)	(0.012243)	(0.0067)
Asinh Citations pre-period	0.1472***	0.133***	0.127***	-0.044	0.134***	0.126***	0.13***	0.0850***
	(0.0081)	(0.00778)	(0.00814)	(0.267031)	(0.007758)	(0.008097)	(0.007189)	(0.0090)
Male	0.069	0.071	0.057	0.915	0.061	0.056	0.076*	0.0542
	(0.0479)	(0.045882)	(0.048169)	(1.385753)	(0.045879)	(0.046558)	(0.037321)	(0.0561)
Early Career Prize	-0.2575	-0.25	-0.272	44.263	-0.243	-0.236	-0.136	-0.2415
	(0.2233)	(0.22288)	(0.229886)	(22.655417)	(0.221762)	(0.234831)	(0.211308)	(0.1872)
Clinical concentration	0.4261***	0.127	0.011	2.085	0.199	0.077	0.124	0.2223
	(0.1306)	(0.116691)	(0.124089)	(3.297669)	(0.114695)	(0.113975)	(0.098198)	(0.1375)
Phd	0.2536***	0.227***	0.227***	1.886	0.246***	0.241***	0.243***	0.5342***
	(0.0577)	(0.053948)	(0.058704)	(2.232738)	(0.054136)	(0.05623)	(0.0556)	(0.0677)
Md	0.2231***	0.131*	0.061	-2.393	0.128*	0.076	0.055	0.0612
	(0.0682)	(0.065454)	(0.072465)	(2.764549)	(0.064826)	(0.067879)	(0.06043)	(0.0777)
Decades after graduation	-0.0003	0.005	0.006	0.018	0.003	0.005	-0.026	-0.1173***
	(0.0281)	(0.028036)	(0.028889)	(1.082879)	(0.028351)	(0.029139)	(0.023233)	(0.0289)
Late Career Recognition	0.3209***	0.196	0.124	9.833*	0.199	0.134	0.281*	0.5451***
	(0.1235)	(0.117401)	(0.127387)	(4.996584)	(0.119125)	(0.123271)	(0.129572)	(0.1023)
Ivy League Undergrade	0.0733	0.04	0.04	0.314	0.045	0.03	0.032	0.1746***
	(0.0588)	(0.055837)	(0.058034)	(1.438286)	(0.056435)	(0.058121)	(0.051373)	(0.0645)
Betweenness centrality								-0.0003
								(0.0003)
Closeness centrality								1.6903
								(2.0470)
Cragg-Donald Wald F statistic	-	33.8552	25.8859	-	40.4088	43.4486	-	44.3189
OIR test p-value	-	0.7824	0.9999	0.9718	0.9906	1	1	1
Community fixed-effect	yes	yes	yes	yes	yes	yes	yes	yes
Contextual variables	no	yes	yes	yes	no	no	no	yes

Notes: Estimates are derived from three models: 2SLS-1, which treats the adjacency matrix $\mathbf{G}$ as exogenous; 2SLS-2, which uses the predicted adjacency matrix $\hat{\mathbf{G}}$ as an instrument for the observed adjacency matrix $\mathbf{G}$; and GMM, which adds a quadratic moment to the instrument used in 2SLS-2. For columns 1-7, the dependent variable is the predicted number of citations, excluding prestige factors. In the last column, the dependent variable is the total number of citations over the study period, weighted by journal impact factor. The following dummy variables are included: Male, Early Career Prize, PhD, MD, Late Career Recognition, and Ivy League Undergrade. The Clinical Concentration variable represents the proportion of clinical papers published by the author during the period. Author's Experience is captured by the variable Decades after Graduation, based on the author's last degree (MD or PhD). The past production is measured by the inverse hyperbolic sine (asinh) of previous citations, weighted by the impact factor of citing journals. Heteroscedasticity-robust standard errors are reported in parentheses. Statistical significance is indicated by $*$, $**$, and $***$ for the 10%, 5%, and 1% levels, respectively.

5.3 Centrality measure comparison

It is useful to compare the model's predictions with those from standard centrality measures not supported by the model. This helps assess how the analysis would change if these variables were included in the estimation. Network centrality measures use different criteria to rank a scientist's importance in the network. As a result, some centrality measures may robustly predict outcomes based on a scientist's position, while others may not be as effective.

Among the most commonly used network centrality measures, closeness and betweenness centrality assess a scientist's position within a network using different criteria. Closeness centrality gauges how quickly a researcher can access or spread information, emphasizing proximity to all other nodes. Betweenness centrality, as defined by Freeman et al. [2002], measures the extent to which a scientist serves as a bridge within the network, based on the probability that they lie on the shortest path between two other nodes. Scientists with high betweenness centrality connect otherwise distant groups, granting them access to diverse information and ideas, which can enhance their productivity and influence. As such, betweenness centrality is particularly relevant in networks where information flow is key. If the influence of these centrality measures on research output were significant, it would suggest that the results reflect information contagion rather than strategic complementarities. This distinction is crucial for understanding whether peer effects in academic networks stem from strategic complementarities or simply from the flow of information. The estimates in the last column of Table (5) support the strategic complementarity argument while rejecting the information contagion hypothesis.

6 Conclusion

In this paper, I presented a theory of scientific peer effects to study knowledge production when authors care about the behaviour of their coauthors. The theory predicts that author's output increase in their Katz-Bonacich centralities, a standard measure of centrality in networks.

I then tested the predictions of our model using a very detailed and unique dataset of scientific researchers in the field of cancer research. I explored the role of network location for peer effects in knowledge production in this field. I then demonstrated that, after controlling for observable characteristics and unobservable network specific factors, an author's position in a network (as measured by her Katz – Bonacich centrality) is a key determinant of their productivity level.

My results validate the importance of peer effects from scientists and not only superstars scientists, producing new evidence that the network position plays an important role in the productivity of scientists. From a general perspective, these results are indicative, in both significance and magnitude, of the relevance of network externalities in scientific productivity. Uncovering the influence of immediate peers and, more generally, the network structure on productivity has important implications, not just for the creation of scientific knowledge, but also for long-run growth potential. The role of peer learning has been featured in

the endogenous growth literature [Romer, 1990]. My results gives an empirical argument toward the roots of peer effects, by emphasizing the critical role of the Katz-Bonacich centrality to explain the productivity of scientists. By controlling for the so-called Matthew effect, this also give new evidence in favor of truly scientific content based peer effects instead of exclusively prestige driven effects. Indeed, the Nash Katz-Bonacich linkage has significant policy implications [Ballester et al., 2006], particularly regarding the optimal network structure when the policymaker aims to maximize total knowledge production [Belhaj et al., 2016].
There is number of possible extensions of this work. First, It would be interesting to unpack the heterogeneity of the peers effects not only based on the network but also making them a function of idiosyncratic characteristics. Second, in this work, we consider a utility function where peer effects are local aggregates, so that it is the sum of peer production that affects utility. It would be interesting to consider a local average instead, where it is the average peer production that matters instead of the aggregate production of coauthors. In that case, we would study the role of conformism in shaping scientific productivity. Finally, output other than published papers could be studied. It has indeed been well documented that social networks are important in the innovation process. It would be interesting to investigate whether the location of an inventor in a network of co-inventors, as measured by their Katz–Bonacich centrality, also has a crucial impact on the individual's patenting level.

A Cancer related papers

To identify cancer-relevant papers, I leverage MeSH (Medical Subject Headings) terms, a hierarchical controlled vocabulary thesaurus managed by the National Library of Medicine (NLM). Professional indexers at the NLM assign MeSH terms to biomedical publications following established protocols, ensuring consistency and context within the entire collection of articles. Significantly, authors of the publications do not participate in selecting MeSH terms. This ensures that the indexing process remains objective, as it is carried out by trained professionals who mitigate the inherent subjectivity of indexing tasks.

MeSH terms serve as a critical tool in biomedical research, enabling researchers to effectively search and categorize vast amounts of literature. By organizing articles into a structured hierarchy, MeSH terms facilitate precise and efficient retrieval of information, which is crucial for advancing scientific knowledge and improving clinical practice. The hierarchical structure of MeSH terms allows for the grouping of related concepts, making it easier to identify overarching themes and specific details within the literature.

The figure below illustrates the MeSH tree structure for cancer-related terms, showcasing how specific terms are nested under broader categories. This visual representation helps in understanding the relationships and hierarchies among different MeSH terms, highlighting their interconnectedness and relevance to cancer research.

Cancer-related papers were identified primarily using the child of the MeSH code C04, i.e., "Neoplasms," which represents the most comprehensive family of MeSH terms for cancer. This category encompasses a

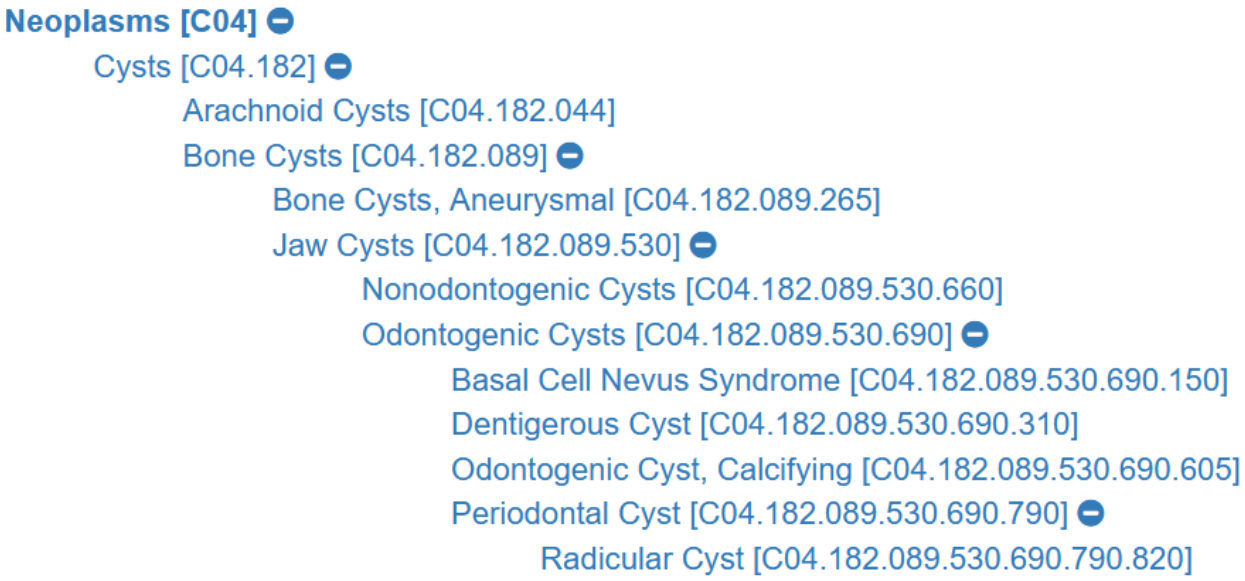

Figure 1: Illustration of the MeSH tree structure for Neoplasms child nodes.

wide range of malignancies, including solid tumors and hematologic cancers, providing a broad foundation for cancer research. In addition to C04, I considered the advice of actual oncologists to include children of the following MeSH terms, which are relevant to various aspects of cancer biology, treatment, and research:

- **Drug Screening Assays, Antitumor** (E05.337.550.200): These assays are crucial for identifying and developing new antitumor agents by testing their efficacy in inhibiting cancer cell growth.
- **Cancer Vaccines** (D20.215.894.200): Vaccines designed to prevent or treat cancer by stimulating the immune system to target cancer cells.
- **Neoplasms** (C04): A broad category encompassing all types of tumors, both benign and malignant.
- **DNA, Neoplasm** (D13.444.308.425): Refers to the genetic material specific to cancer cells, which can be targeted for diagnosis or therapy.
- **Drug Resistance, Neoplasm** (G07.690.773.984.395): The ability of cancer cells to resist the effects of chemotherapy, posing a significant challenge in cancer treatment.
- **Neoplasm Proteins** (D12.776.624): Proteins specifically associated with tumors, which can serve as biomarkers or therapeutic targets.
- **Biomarkers, Tumor** (D23.101.140): Biological molecules found in blood, other body fluids, or tissues that indicate the presence of cancer.
- **Antigens, Neoplasm** (D23.050.285): Substances produced by tumor cells that can trigger an immune response.
- **Oncogenic Viruses** (B04.613): Viruses that can cause cancer by integrating their genetic material into the host genome.

- **Tumor Cells, Cultured** (A11.251.860): Cancer cells grown in laboratory conditions for research purposes.
- **Neoplasm Proteins** (D12.776.624): Proteins associated with tumors, important for understanding cancer biology and developing treatments.
- **Chemotherapy, Cancer, Regional Perfusion** (E04.292.425): A technique to deliver high doses of chemotherapy directly to the tumor site.
- **Antineoplastic Agents** (D27.505.954.248): Drugs used to treat cancer by inhibiting the growth of malignant cells.
- **Receptors, Tumor Necrosis Factor** (D12.776.543.750.705.852.760): Receptors involved in the signaling pathways that can lead to tumor cell death.
- **Tumor Escape** (G12.900): Mechanisms by which cancer cells evade the immune system.
- **Neoplastic Stem Cells** (A11.872.650): Stem cells within tumors that have the ability to self-renew and drive cancer progression.
- **Carcinogens** (D27.888.569.100): Substances capable of causing cancer in living tissue.
- **Gammaretrovirus** (B04.820.650.375): A type of virus that can insert its genetic material into the host genome, potentially causing cancer.
- **Antibodies, Neoplasm** (D12.776.377.715.548.114.240): Antibodies used to target and neutralize cancer cells.
- **Receptors, Immunologic** (D12.776.543.750.705): Receptors on immune cells that can be manipulated to enhance the immune response against cancer.
- **Tumor Necrosis Factors** (D23.529.374.750): Proteins involved in the destruction of cancer cells.
- **Biomarkers, Tumor** (D23.101.140): Indicators used to detect cancer or monitor its progression.
- **Radiotherapy** (E02.815): The use of high-energy radiation to kill or shrink cancer cells.

B Computing Cognitive proximity

To measure the distance, or rather proximity, in intellectual or "ideas space" between pairs of scientists, I constructed a variable based on MeSH terms. Delineating the boundaries of scientific fields is challenging, as most research can be classified in numerous ways, and consensus among scientists on specific categorizations is often lacking. Traditional measures based on shared department affiliation or broad scientific field distinctions (e.g., cell vs. molecular biology) are inadequate for this purpose. Instead of positioning scientists at fixed points in ideas space, this method provides a cost-effective and convenient way to measure their relative positions.

MeSH terms offer a fine-grained level of detail, making them particularly useful for this task. For example, a paper motivating the 2018 Nobel Prize in Medicine [Ishida et al., 1992] is labeled with 27 distinct descriptors, ranging from general terms like "Animals" and "Humans" to specific ones like "CD3 Complex/Genetics" and "T Cells/Immunology." However, during the construction of MeSH vectors for 2001-2005, I found that such detailed categories could result in overly specific and uninformative comparisons across scientists. Fine-grained categorizations might obscure meaningful patterns by focusing too narrowly on minor differences. To ensure more informative and comparative analysis, I opted for a coarser categorization of MeSH terms, utilizing second-level categories. For instance, the MeSH term "Receptors, Immunologic," originally categorized under the tree number D12.776.543.750.705, was grouped under the broader category D12.776. I then constructed a vector for each author, where each entry is set to 1 if the author has at least one paper with a MeSH term belonging to the specified categories over the period, and 0 otherwise. This approach allowed us to capture broader trends and more meaningful comparisons between scientists' intellectual proximity. The cosine index [Breschi et al., 2003] was used to calculate the extent of overlap between the author's research interest and each of their co-authors' interests. Here, it was assumed that the higher the overlap between two authors in terms of the breadth and depth of their research interest, the closer they are in the cognitive space. The cosine index between author i and j, which is used as the independent variable $CognProx_{ij}$, is calculated as follows:

$$CognProx_{ij} = \frac{\sum_k a_{ik} a_{jk}}{\sqrt{\sum_k a_{ik}^2} \sqrt{\sum_k a_{jk}^2}}$$

where a_{ik} refers to the presence of categorie k mesh in all the papers published by author i. Obviously, $CognProx_{ij} = 1$ indicates that the two authors are exactly the same in terms of their research interest, and if there is no common research interest between the two authors, $CognProx_{ij} = 0$. Therefore, high cosine values indicate increased overlap between the knowledge bases of two authors, in terms of their similarity.

C Robustness : Alternatives time period

References

Rajshree Agarwal, Alexander Oettl, and Scott Stern. The impact of academic entrepreneurship on university research performance: Evidence from patenting at u.s. universities. *Research Policy*, 44(7):1288–1305, 2015.

Ajay Agrawal, John McHale, and Alexander Oettl. How stars matter: Recruiting and peer effects in evolutionary biology. *Research Policy*, 46(4):853–867, 2017.

Pierre Azoulay, Joshua S Graff Zivin, and Jialan Wang. Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589, 2010.

Table 6: Results for the period 2006-2010

	Dependent variable:						
	Weighted number of citations						
	OLS	2SLS-1	2SLS-2	GMM	2SLS-1	2SLS-2	GMM
λ		0.04***	0.056***	0.053**	0.041***	0.055***	0.055***
		(0.003516)	(0.005756)	(0.019665)	(0.003569)	(0.005354)	(0.013227)
Asinh Citations pre-period	0.1668***	0.144***	0.139***	0.146***	0.144***	0.136***	0.141***
	(0.0173)	(0.015783)	(0.017889)	(0.024652)	(0.015862)	(0.016598)	(0.01162)
Male	0.1441*	0.097	0.079	0.099	0.116	0.106	0.099
	(0.0851)	(0.081452)	(0.087381)	(0.100499)	(0.081215)	(0.082241)	(0.051252)
Early Career Prize	-0.572	-0.702*	-0.648	-0.074	-0.668*	-0.701*	-0.853*
	(0.4167)	(0.321413)	(0.357286)	(0.657657)	(0.332825)	(0.351499)	(0.382023)
Clinical concentration	1.1457***	0.6*	0.367	0.504	0.62**	0.441	0.466**
	(0.2696)	(0.242636)	(0.263058)	(0.372879)	(0.236993)	(0.238302)	(0.142199)
Phd	0.243**	0.238*	0.215	-0.098	0.293**	0.31**	0.343***
	(0.1136)	(0.104574)	(0.110945)	(0.171646)	(0.10298)	(0.103559)	(0.071854)
Md	0.1174	0.106	0.148	0.014	0.02	-0.014	-0.038
	(0.1244)	(0.116013)	(0.129585)	(0.184777)	(0.111585)	(0.111399)	(0.079778)
Decades after graduation	-0.2879***	-0.255***	-0.272***	-0.248***	-0.263***	-0.255***	-0.273***
	(0.058)	(0.054103)	(0.06174)	(0.07101)	(0.05366)	(0.054743)	(0.038405)
Late Career Recognition	0.7982***	0.581**	0.522**	0.317	0.623***	0.563**	0.597**
	(0.2013)	(0.183843)	(0.1879)	(0.257383)	(0.181218)	(0.185141)	(0.200646)
Ivy League Undergrade	-0.0464	-0.069	-0.094	-0.195	-0.055	-0.058	-0.01
	(0.1006)	(0.090695)	(0.097996)	(0.122729)	(0.091868)	(0.093783)	(0.06744)
Cragg-Donald Wald F statistic	-	15.5498	9.1578	-	17.6103	14.4636	-
OIR test p-value	-	0.7226	0.994	1	0.9905	0.9959	1
Community fixed-effect	yes	yes	yes	yes	yes	yes	yes
Contextual variables	no	yes	yes	yes	no	no	no

Note: Heteroscedasticity-robust standard errors in parentheses.
***p<0.01, **p<0.05, *p<0.1

Pierre Azoulay, Toby Stuart, and Yanbo Wang. Matthew: Effect or fable? *Management Science*, 60(1): 92–109, 2014.

Pierre Azoulay, Christian Fons-Rosen, and Joshua S Graff Zivin. Does science advance one funeral at a time? *American Economic Review*, 109(8):2889–2920, 2019.

Coralio Ballester, Antoni Calvó-Armengol, and Yves Zenou. Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417, 2006.

Michael Balzer and Adhen Benlahlou. Mitigating consequences of prestige in citations of publications. *arXiv preprint arXiv:2411.05584*, 2024.

Mohamed Belhaj, Sebastian Bervoets, and Frédéric Deroïan. Efficient networks in games with local complementarities. *Theoretical Economics*, 11(1):357–380, 2016.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Phillip Bonacich. Power and centrality: A family of measures. *American journal of sociology*, 92(5):1170–1182, 1987.

Table 7: Results for the period 2015-2018

	Dependent variable:						
	Weighted number of citations						
	OLS	2SLS-1	2SLS-2	GMM	2SLS-1	2SLS-2	GMM
λ		0.023***	0.034***	0.031**	0.023***	0.032***	0.031***
		(0.001542)	(0.002192)	(0.010867)	(0.001532)	(0.002083)	(0.006286)
Asinh Citations pre-period	0.1923***	0.162***	0.143***	0.134***	0.163***	0.153***	0.151***
	(0.0132)	(0.012178)	(0.013084)	(0.018523)	(0.012114)	(0.012362)	(0.009314)
Male	0.1668***	0.125*	0.163*	0.215**	0.135*	0.124*	0.161***
	(0.0643)	(0.058727)	(0.064135)	(0.083005)	(0.058151)	(0.058593)	(0.035041)
Early Career Prize	-0.0445	0.012	0.013	-0.256	0.026	0.052	0.12
	(0.2398)	(0.214384)	(0.218786)	(0.326095)	(0.214455)	(0.211051)	(0.1439)
Clinical concentration	0.6811***	0.376*	0.2	-0.027	0.395*	0.291	0.374***
	(0.1726)	(0.162099)	(0.176654)	(0.180671)	(0.160078)	(0.160291)	(0.07898)
Phd	0.1197	0.08	0.046	-0.099	0.102	0.095	0.117*
	(0.0885)	(0.081691)	(0.088837)	(0.128626)	(0.081643)	(0.082174)	(0.057942)
Md	0.1529	-0.005	-0.055	0.118	-0.045	-0.117	-0.163**
	(0.0943)	(0.089223)	(0.101179)	(0.142389)	(0.086837)	(0.088387)	(0.059983)
Decades after graduation	-0.3833***	-0.335***	-0.305***	-0.295***	-0.335***	-0.318***	-0.326***
	(0.0358)	(0.033518)	(0.035604)	(0.052248)	(0.0333)	(0.03392)	(0.027285)
Late Career Recognition	0.6532***	0.537***	0.493**	-0.181	0.533***	0.49***	0.492***
	(0.1486)	(0.140808)	(0.156214)	(0.19726)	(0.140854)	(0.144741)	(0.120878)
Ivy League Undergrade	-0.0871	-0.141	-0.165*	-0.416***	-0.142	-0.162*	-0.133*
	(0.0823)	(0.072945)	(0.077039)	(0.098574)	(0.072991)	(0.074589)	(0.05479)
Cragg-Donald Wald F statistic	-	25.7953	17.4919	-	29.7328	27.4617	-
OIR test p-value	-	0.2177	1	1	0.9963	1	1
Community fixed-effect	yes	yes	yes	yes	yes	yes	yes
Contextual variables	no	yes	yes	yes	no	no	no

Note: Heteroscedasticity-robust standard errors in parentheses.
***p<0.01, **p<0.05, *p<0.1

George J Borjas and Kirk B Doran. Which peers matter? the relative impacts of collaborators, colleagues, and competitors. *The Review of Economics and Statistics*, 97(5):1104–1117, 2015.

Ron Boschma. Proximity and innovation: a critical assessment. *Regional studies*, 39(1):61–74, 2005.

Cathy J Bradley, K Robin Yabroff, Bassam Dahman, Eric J Feuer, Angela Mariotto, and Martin L Brown. Productivity costs of cancer mortality in the united states: 2000–2020. *JNCI: Journal of the National Cancer Institute*, 100(24):1763–1770, 2008.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55, 2009.

Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Peer effects in networks: A survey. *Annual Review of Economics*, 12(1):603–629, 2020.

Stefano Breschi and Francesco Lissoni. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *Journal of economic geography*, 9(4):439–468, 2009.

Stefano Breschi, Francesco Lissoni, and Franco Malerba. Knowledge-relatedness in firm technological diversification. *Research policy*, 32(1):69–87, 2003.

Linus Dahlander and Daniel A McFarland. Ties that last: Tie formation and persistence in research collaborations over time. *Administrative science quarterly*, 58(1):69–110, 2013.

Waverly W Ding, Sharon G Levin, Paula E Stephan, and Anne E Winkler. The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Science*, 56(9): 1439–1461, 2010.

Lorenzo Ductor, Sanjeev Goyal, and Anja Prummer. Gender and collaboration. *Review of Economics and Statistics*, 105(6):1366–1378, 2023.

Marcel Fafchamps, Marco J Van der Leij, and Sanjeev Goyal. Matching and network effects. *Journal of the European Economic Association*, 8(1):203–231, 2010.

Linton C Freeman et al. Centrality in social networks: Conceptual clarification. *Social network: critical concepts in sociology. Londres: Routledge*, 1:238–263, 2002.

Richard B Freeman and Wei Huang. Collaborating with people like me: Ethnic coauthorship within the united states. *Journal of Labor Economics*, 33(S1):S289–S318, 2015.

Richard B Freeman, Ina Ganguli, and Raviv Murciano-Goroff. Why and wherefore of increased scientific collaboration. In *The changing frontier: Rethinking science and innovation policy*, pages 17–48. University of Chicago Press, 2014.

B Ian Hutchins, Kirk L Baker, Matthew T Davis, Mario A Diwersy, Ehsanul Haque, Robert M Harriman, Travis A Hoppe, Stephen A Leicht, Payam Meyer, and George M Santangelo. The nih open citation collection: A public access, broad coverage resource. *PLoS biology*, 17(10):e3000385, 2019.

Yasumasa Ishida, Yasutoshi Agata, Keiichi Shibahara, and Tasuku Honjo. Induced expression of pd-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *The EMBO journal*, 11(11):3887–3895, 1992.

Xavier Jaravel, Neviana Petkova, and Alex Bell. Team-specific capital and innovation. *American Economic Review*, 108(4-5):1034–1073, 2018.

Koen Jochmans. Peer effects and endogenous social interactions. *Journal of Econometrics*, 235(2):1203–1214, 2023.

Benjamin F Jones. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317, 2009.

In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97, 2009.

J Sylvan Katz and Ben R Martin. What is research collaboration? *Research policy*, 26(1):1–18, 1997.

Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

Harry H Kelejian and Gianfranco Piras. Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics*, 46:140–149, 2014.

Harry H Kelejian and Ingmar R Prucha. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of econometrics*, 157(1):53–67, 2010.

Lung-fei Lee. Gmm and 2sls estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514, 2007.

Lung-Fei Lee, Xiaodong Liu, Eleonora Patacchini, and Yves Zenou. Who is the key player? a network analysis of juvenile delinquency. *Journal of Business & Economic Statistics*, 39(3):849–857, 2021.

Xiaodong Liu and Lung-fei Lee. Gmm estimation of social interaction models with centrality. *Journal of Econometrics*, 159(1):99–115, 2010.

Janet C Long, Frances C Cunningham, Peter Carswell, and Jeffrey Braithwaite. Patterns of collaboration in complex networks: the example of a translational research network. *BMC Health Services Research*, 14: 1–10, 2014.

J Loscalzo and AL Barabási. Network science, 2016.

Jacques Mairesse and Laure Turner. Measurement and explanation of the intensity of co-publication in scientific research: An analysis at the laboratory level. *New Frontiers in the Economics of Innovation and New Technology: Essays in Honour of Paul A. David. Edward Elgar, Cheltenham and Northampton*, pages 255–295, 2006.

Charles F Manski. Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542, 1993.

Sadao Nagaoka and Hideo Owan. Author ordering in scientific research: evidence. *Quarterly Journal of Economics*, 109(4):1185–1209, 2014.

Ikujirō Nonaka and Hirotaka Takeuchi. The knowledge-creating company. *Harvard business review*, 85(7/8): 162, 2007.

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

M Polanyi. The tacit dimension, anchor day. *New York*, 1966.

Paul M Romer. Endogenous technological change. *Journal of political Economy*, 98(5, Part 2):S71–S102, 1990.

Neil R Smalheiser, Vetle I Torvik, et al. Author name disambiguation. *Annual review of information science and technology*, 43(1):1, 2009.

Vetle I Torvik, Marc Weeber, Don R Swanson, and Neil R Smalheiser. A probabilistic similarity metric for medline records: A model for author name disambiguation. *Journal of the American Society for information science and technology*, 56(2):140–158, 2005.

Brian Uzzi. The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American sociological review*, pages 674–698, 1996.

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

Fabian Waldinger. Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. *The Review of Economics and Statistics*, 98(5):811–831, 2016.

Stanley Wasserman and Katherine Faust. Social network analysis: Methods and applications. 1994.

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

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

Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F Rousseau, Xin Li, Weijia Xu, Vetle I Torvik, et al. Building a pubmed knowledge graph. *Scientific data*, 7(1):205, 2020.

Harriet A Zuckerman. Patterns of name ordering among authors of scientific papers: A study of social symbolism and its ambiguity. *American Journal of Sociology*, 74(3):276–291, 1968.