On the origin of scientific peer effects, strategic complementarity or conformity?

Adhen Benlahlou[a,1,]

[a]*Bielefeld University, Chair for Economic Theory and Computational Economics, Universitätsstraße 25, 33615, Bielefeld, Germany*

Abstract

Collaboration plays a critical role in scientific productivity, yet the mechanisms through which peer effects drive innovation remain underexplored. This paper investigates the microfoundations of scientific peer effects, focusing on how collaboration influences individual productivity. While existing literature emphasizes the importance of peer effects, it has not fully addressed the underlying processes. This work contributes by offering a clearer understanding of these mechanisms and providing insights for innovation policy, particularly in contexts where collaboration drives knowledge creation. The results show that peer effects arise from strategic complementarity, where collaborating with more productive coauthors enhances individual performance. These findings underscore the significant role of network externalities in shaping scientific outcomes and the structure of scientific collaboration.

Keywords: productivity, innovation, network, peer effects, knowledge production, scientific collaboration

JEL: C72, D85, D43, L14, Z13

1. Introduction

Technological advancement is a key driver of economic growth, but the mechanisms behind knowledge creation remain complex. Endogenous growth models suggest that new knowledge arises from recombining existing knowledge stocks (Romer, 1990; Weitzman, 1998), though the exact processes are not fully understood. A growing body of research highlights the critical role of scientists' individual characteristics—such as curiosity, experience, and expertise—in driving innovation (Fleming, 2001), with these traits varying significantly among individuals and influencing productivity (Fortunato et al., 2018). However, increasing evidence shows that knowledge creation is increasingly collaborative (Wuchty et al., 2007). Collaboration allows scientists to exchange ideas and tap into a wider knowledge base, enhancing productivity. Studies by Azoulay et al. (2010) and Waldinger (2012) have explored the role of peer interactions and knowledge spillovers in academia, using quasi-natural experiments to estimate causal effects and highlight the importance of these coauthorship network in advancing

scientific progress.

Social networks research has shown that an agent's position within a network correlates with their economic performance. This relationship is often analyzed using centrality measures, which quantify an agent's importance based on their network connections. Benlahlou (2024) demonstrate that Katz-Bonacich centrality outperforms other centrality measures in capturing the link between network position and scientific productivity. However, Katz-Bonacich centrality can be interpreted through two distinct microfoundations: strategic complementarity and social conformity (Ushchev and Zenou, 2020). While both mechanisms are represented in the same Nash Katz-Bonacich linkage, they reflect different processes and lead to distinct policy implications. Therefore, disentangling these mechanisms is crucial for understanding their roles in scientific productivity and for developing more effective network-based policies that promote knowledge creation.

Understanding the magnitude and sources of individual-level spillovers is crucial for innovation policy. The impact of scientists' actions extends beyond their own behavior, generating broader "multiplier effects" that shape the overall production process. By exploring these microfoundations, I aim to deepen our understanding of how scientists' interactions and decisions influence scientific productivity. Combining a micro-founded model with a robust empirical strategy allows me to identify causal relationships, advancing our theoretical understanding of knowledge production and offering insights for more effective policies to foster scientific progress and economic growth.

The microfoundations of this model build on the framework developed by Liu et al. (2014), which integrates two key models of peer effects in networks: the local-aggregate and local-average models. In the local-aggregate model (Ballester et al., 2006; Belhaj et al., 2016; Calvó-Armengol et al., 2009; König et al., 2014), peer effects arise from the total output of coauthors. Here, the productivity of a scientist's collaborators directly increases their marginal utility from producing science. Conversely, the local-average model (Patacchini and Zenou, 2012; Boucher et al., 2014) treats peer effects as a social norm. In this model, scientists incur a cost for deviating from the average production level of their coauthors, leading them to adjust their output to align with the social norm defined by their peers' average productivity.

I test this model exploiting a unique dataset focused on cancer-related research, a key innovation sector rooted in scientific advances. These advances stem from the voluntary sharing of knowledge through collaboration rather than individual efforts. The relevance of this focus is underscored by the significant economic impact of cancer. According to the National Institutes of Health, cancer care cost

the US $147.5 billion in 2015, and lost productivity due to early deaths from cancer resulted in an additional $134.8 billion loss in 2005. This dataset was compiled by extensively leveraging information from CVs gathered from various university websites and specialized professional social networks. The résumés were supplemented with disambiguated publication data from the PubMed Knowledge Graph, along with additional metadata constructed by the Torvik research group.

I adopt an instrumental variable (IV) strategy to estimate the best-response function implied by the theoretical model, allowing me to identify peer effects while distinguishing them from contextual and correlated effects. To address the potential endogeneity of collaboration networks, I use predicted collaboration networks based on predetermined dyadic characteristics to construct IVs and identify the causal effect of collaboration (Liu et al., 2014). Additionally, following Jochmans (2023), I use average characteristics of indirect coauthors to construct IVs for the average characteristics of coauthors.

Exploiting a clean identification, I find that peer effects arise from strategic complementarity, even if estimation in isolation of a conformity model may lead to incorrect conclusions. These findings illustrate the potential significance and magnitude of network externalities in scientific productivity and provide insights into the mechanisms underlying peer effects. Unpacking the origin of these peer effects is particularly beneficial. For instance, ruling out the conformity mechanism and providing support for the strategic complementarity mechanism has important implications for both comparative statics and the design of optimal network policies.

The remainder of the paper is organized as follows. In Section 1, I provide a theoretical perspective on scientific peer effects that guides the subsequent empirical analysis. In Section 2, I discuss the identification threats and outline the methodology used to estimate the model. Section 3 describes the data used in the analysis. Section 4 presents the results from a structural estimation to estimate peer effects among scientists, allowing for heterogeneity with respect to network position. Section 5 concludes.

2. Theoretical framework

The main empirical question I address is whether peer effects in scientific production stem from strategic complementarity and/or conformity. To test this hypothesis, I use the peer effects model developed by Liu et al. (2014), which is well-suited for the research environment. This model captures both strategic complementarity and conformism, while also encompassing two standard models of peer effects: the local-average and local-aggregate models.

The model rests on two key assumptions: first, that scientists benefit from the increased production of their peers; and second, that they incur a cost when deviating from the norms of their reference group. In equilibrium, a scientist's production is a function of their network position, with their incentives shaped by the choices of coauthors through both strategic complementarity and conformism, as well as by their own individual scientific productivity.

2.1. The Network and Notations

Suppose a finite set of scientists $N = \{1, \ldots, n\}$ is divided into communities, where $N_c = \{1, \ldots, n_c\}$ represents the scientists in the c-th community. Coauthorship connections within community c are tracked through its adjacency matrix $G_c = [g_{ij,c}]$, where $g_{ij,c} = 1$ if i and j are coauthors, and $g_{ij,c} = 0$ otherwise. Additionally, $g_{ii,c} = 0$. The reference group of scientist i in community c consists of i's coauthors, denoted as $N_{i,c} = \{j \neq i \mid g_{ij,c} = 1\}$. The size of $N_{i,c}$, termed the degree of i in graph theory, is $|N_{i,c}| = \{j \neq i \mid g_{ij,c} = 1\}|$. Let $G_c^* = [g_{ij,c}^\star]$, where $g_{ij,c}^\star = g_{ij,c}/g_{i,c}$.

I denote by $y_{i,c}$ the production level of researcher i and by the population production profile of community c.

2.2. Modeling Benefits of Collaboration

The literature on peer effects in scientific productivity implicitly recognizes the concept of strategic complementarity, where the actions of peers positively influence a scientist's productivity and the quality of their research. These effects manifest through various means, including the sharing of insights, feedback provision, and joint development of methodologies. Collaborative efforts, such as co-authorship, combine diverse expertise and foster an environment rich in intellectual exchange.
Working with innovative and productive peers often grants access to valuable ideas, especially when strong connections are maintained. Early exposure to such ideas significantly boosts productivity, suggesting that scientists closely associated with highly productive peers tend to exhibit greater output. Thus, an author's productivity is influenced not only by their individual traits but also by the productivity of their co-authors, illustrating the recursive nature of productivity and its dependence on peer interactions.
The scientific peer effects litterature highlights two key factors that affect the benefits of collaboration: strategic complementarity and heterogeneity in production returns. To capture these dynamics, I model productivity as a function of individual output, personal productivity, and the output of

co-authors:

$$\underbrace{\left(\pi^{\star}_{i,c} + \lambda_1 \sum_{j=1}^{n_c} g_{ij,c} y_{j,c}\right) y_{i,c}}_{\text{benefits}}$$

In this model, $y_{i,c}$ represents the production level of scientist i, and the term $\left(\pi^{\star}_{i,c} + \lambda_1 \sum_{j=1}^{n_c} g_{ij,c} y_{j,c}\right)$ represents the return to production. Here, $\pi^{\star}_{i,c}$ captures individual heterogeneity in production returns, and $\sum_{j \in N_{i,c}} y_{j,c}$ represents the aggregate production of i's co-authors. The social multiplier coefficient, $\lambda_1 \geq 0$, reflects the influence of peer production on individual output. Importantly, this model allows for heterogeneity in the network structure, as the aggregate production of co-authors differs for each scientist, even if all co-authors choose the same production level.

2.3. Modeling the Cost of Collaboration

In the sociology of science, recognition plays a crucial role in shaping scientists' behaviors and motivations. A substantial body of literature emphasizes that scientists seek validation and recognition from their peers, often valuing these over obscurity. This recognition serves as an affirmation of the quality and significance of their work (Zuckerman, 1977; Latour and Woolgar, 1986). However, the pursuit of recognition and validation within collaborative environments incurs various costs for individual scientists.

In team-based research endeavors, recognizing individual contributions becomes more complex as the lines between individual and collective achievements blur. The subjective nature of recognition and validation within scientific communities adds another layer of complexity to the evaluation process. Researchers may have differing perceptions of the significance and impact of their work, leading to potential conflicts and dissatisfaction. Given these challenges, reassurance through recognition becomes increasingly important in the social dynamics of knowledge production. Despite their expertise, scientists often lack absolute certainty about the worth and impact of their work, making peer recognition a vital form of reassurance and validation that affirms the value of their contributions within the broader scientific community (Merton, 1973).

Individuals typically compare themselves to those they perceive as having similar attitudes or abilities, often disregarding those who are markedly different (Festinger, 1954). Peer groups, consisting of individuals of similar rank and ability, serve as essential reference points for self-evaluation. These groups provide a realistic benchmark for assessing one's position and achievements within a broader social context. Peer groups act as standards or checkpoints to evaluate situations and one's position within

them (Mas and Moretti, 2009). The desire to relate to and be accepted by a group drives ambition and competitive behavior, as individuals seek alignment with their peers (Shibutani, 1955).
To account for the costs incurred by a researcher, I follow the standard approach in economics for modeling conformity, i.e., using an increasing function of the distance to the average output of peers. However, it is interesting to note that, in my case, the social norm is not exogenously determined. In addition to the cost induced by a deviation from the average output of the co-authors, I introduce a quadratic term to capture the increased costs associated with a larger production.

$$\underbrace{\frac{1}{2}\left[y_{i,c}^2 + \lambda_2(y_{i,c} - \sum_{j=1}^{n_c} g_{ij,c}^{\star} y_{j,c})^2\right]}_{\text{cost}}$$

Here, $\lambda_2 \geq 0$ captures the intensity of the conformity effect.

2.4. The Peer Effects Game

Given the underlying network structure represented by the adjacency matrix G_c, individuals in community c simultaneously decide how much science they produce $y_{i,c}$ to maximize the following utility function:

$$\begin{aligned} u_{i,c}(y_{i,c}) \equiv u_{i,c}\left(y_{i,c}, \mathbf{Y}_c, \mathbf{G}_c\right) = & \underbrace{\left(\pi_{i,c}^{\star} + \lambda_1 \sum_{j=1}^{n} g_{ij,c} y_{j,c}\right) y_{i,c}}_{\text{benefits}} \\ & - \underbrace{\frac{1}{2}\left[y_{i,c}^2 + \lambda_2(y_{i,c} - \sum_{j=1}^{n} g_{ij,c}^{\star} y_{j,c})^2\right]}_{\text{cost}}. \end{aligned} \tag{1}$$

Now, I will point out some important properties of the utility function (1), which provide useful intuition about the mechanism at work. First, when scientist j produces, she exerts a negative externality on her coauthor i if and only if the production of i differs from i's social norm. Second, productions are strategic complements, meaning that the higher the production of a scientist's coauthor, the higher the individual's marginal utility of producing science. Hence, a scientist's utility is positively influenced by the collective production of her coauthors and adversely affected by the deviation from the average production of her coauthors. Deriving from the first-order condition of utility maximization, the best-reply function of author i is then expressed as:

$$y_{i,c} = \phi_1 \sum_{j=1}^{n_c} g_{ij,c} y_{j,c} + \phi_2 \sum_{j=1}^{n_c} g^{\star}_{ij,c} y_{j,c} + \pi_{i,c} \tag{2}$$

where $\phi_1 = \frac{\lambda_1}{1+\lambda_2}$, $\phi_2 = \frac{\lambda_2}{1+\lambda_2}$, and $\pi_{i,c} = \frac{\pi^{\star}_{i,c}}{1+\lambda_2}$. As $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, I have $\phi_1 \geq 0$ and $0 \leq \phi_2 < 1$. The coefficient ϕ_1 is called the local-aggregate endogenous peer effect. As $\phi_1 \geq 0$, this coefficient reflects strategic complementarity in efforts. The coefficient ϕ_2 is called the local-average endogenous peer effect, which captures the taste for conformity. Note that $\frac{\phi_1}{\phi_2} = \frac{\lambda_1}{\lambda_2}$. That is, the relative magnitude of ϕ_1 and ϕ_2 is the same that of the strategic complementarity coefficient λ_1 and the social-conformity coefficient λ_2.

I denote by $g_c^{\max}$ the highest degree in network c, i.e. $g_c^{\max} = \max_i g_{i,c}$. Let $\Pi_c = (\pi_{1,c}), \cdots, \pi_{nc,c}$. The Nash equilibrium of the general network model is characterized by the following proposition:

Proposition 1 (Liu et al., 2014)

If $\phi_1 \geq 0$, $\phi_2 \geq 0$, and $g_c^{\max}\lambda_1\lambda_2 < 1$, then the network game with payoffs (1) has a unique interior Nash equilibrium in pure strategies given by:

$$\mathbf{Y}_c = \left(\mathbf{I}_{nc} - \phi_1 \mathbf{G}_c - \phi_2 \mathbf{G}_c^{\star}\right)^{-1} \Pi_c$$

I am dealing with a game with strategic complementarities, so the higher the production of coauthors, the higher my marginal utility of raising my own production. Consequently, there is a problem of equilibrium existence since there is no bound on the production of each agent. Since ϕ_1 and ϕ_2 express the intensity of these complementarities, the condition $g_c^{\max}\phi_1\phi_2 < 1$ guarantees the existence of equilibrium by bounding the relative strength of strategic complementarities of production with respect to the amplifying power of the network. The linear best-reply functions combined with strategic complementarities rule out the multiple equilibria case.

2.5. Model Discussion

In the pure conformity scenario (local-average model, i.e., $\lambda_1 = 0$), a scientist's utility is influenced by how much her production deviates from the average production of her coauthor group. Therefore, the closer a scientist's production is to the average production of her coauthors, the higher her equilibrium utility will be. In contrast, in the pure strategic complementarity scenario (local-aggregate model, i.e., $\lambda_2 = 0$), a scientist's utility is impacted by the total production of the reference group. Thus, the more active coauthors a scientist has, the higher her equilibrium utility. In the local-average model,

network positions do not affect equilibrium effort if all scientists are ex ante identical. However, in the pure strategic complementarity scenario, different network positions result in different equilibrium production levels, even if scientists are ex ante identical in terms of individual productivity.

These two mechanisms have fundamentally different equilibrium implications. When a scientist's utility is influenced by the aggregate production (strategic complementarities) of coauthors, their position in the network affects their equilibrium production. Conversely, when a deviation from the social norm is costly, all scientists will conform to the production level of their reference group, resulting in uniform production if they are ex ante identical.

Each of the two mechanisms – strategic complementarities, and conformity – is capable of generating the same empirical phenomenon: Katz-Bonacich linkage to the equilibrium production. Consequently, distinguishing the mechanism is difficult.

However, this distinction implies that the policy implications of the two mechanisms are quite different. In the case of conformity, the only way to influence scientists' choices and outcomes is by altering the social norm of the community. This means that the policy must affect most scientists in the community to be effective. Therefore, group-based policies, such as topic-based initiatives, are suitable for this model. Conversely, with strategic complementarities, due to social multiplier effects, targeting a single scientist can still yield positive outcomes, as that scientist will influence her coauthors. Hence, scientist-based policies, such as targeted subsidies, can be effectively implemented in this scenario.

3. Identification of the econometric network model

The following discussion builds upon the foundational work by Liu et al. (2014), extending their analysis to account for endogenous interactions within the network and heteroscedastic innovations.

3.1. Econometric network model

The specification of the econometric model follows the equilibrium best-reply function of the network game so that it has a clear microfoundation. Let the ex ante heterogeneity $\pi_{i,c}$ of scientist i in community c be

$$\pi_{i,c} = \eta_c + x_{i,c}^{\intercal}\beta_1 + \sum_{j=1}^{n_c} g_{ij,c}^{\star} x_{j,c}^{\intercal}\beta_2 + u_{i,c},$$

where $x_{i,c}$ is a p-dimensional vector of exogenous variables, η_c a constant term, β_1 is a vector of suitable dimension, where the error term u is assumed to be independently distributed but allowed to

be heteroscedastic with $\Sigma \equiv \mathbb{E}(\mathbf{u}\mathbf{u}^{\intercal}) = \sigma_i^2$ being a diagonal matrix., and β_1, β_2 are corresponding parameters. From the best-reply function (2), the general econometric network model is

$$y_{i,c} = \phi_1 \sum_{j=1}^{n_c} g_{ij,c} y_{j,c} + \phi_2 \sum_{j=1}^{n_c} g^{\star}_{ij,c} y_{j,c} + \eta_c + x^{\intercal}_{i,c}\beta_1 + \sum_{j=1}^{n_c} g^{\star}_{ij,c} x^{\intercal}_{j,c}\beta_2 + u_{i,c}, \tag{3}$$

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

$$\mathbf{Y}_c = \phi_1 \mathbf{G}_c \mathbf{Y}_c + \phi_2 \mathbf{G}^{\star}_c \mathbf{Y}_c + \eta_c \iota_{nc} + \mathbf{X}_c \beta_1 + \mathbf{G}^{\star}_c \mathbf{X}_c \beta_2 + \mathbf{u}_c.$$

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 $\mathbf{Y} = \left(\mathbf{Y}_1^{'}, \cdots, \mathbf{Y}_{\bar{c}}^{'}\right)^{'}$, $\mathbf{X} = \left(\mathbf{X}_1^{'}, \cdots, \mathbf{X}_{\bar{c}}^{'}\right)^{'}$, $\mathbf{u} = \left(\mathbf{u}_1^{'}, \cdots, \mathbf{u}_{\bar{c}}^{'}\right)^{'}$, $\eta = \left(\eta_1^{'}, \cdots, \eta_{\bar{c}}^{'}\right)^{'}$, $\mathbf{G} = \{G_c\}_{c=1}^{c=\bar{c}}$, $\mathbf{G} = \{G_c\}_{c=1}^{c=\bar{c}}$ and $\mathbf{L} = \{\iota_{n,c}\}_{c=1}^{c=\bar{c}}$. The econometric model can be written a

$$\mathbf{Y} = \phi_1 \mathbf{G}\mathbf{Y} + \phi_2 \mathbf{G}^{\star}\mathbf{Y} + \mathbf{X}_c\beta_1 + \mathbf{G}^{\star}\mathbf{X}\beta_2 + \mathbf{L}\eta + \mathbf{u}$$

The network-specific parameters, denoted by η, are allowed to vary with $\mathbf{G}$, $\mathbf{G}^*$, and $\mathbf{X}$, similar to the structure in a fixed-effects panel data model. To mitigate the incidental parameter problem that arises when the number of groups is large, we eliminate the term $\mathbf{L}\eta$ by applying the deviation from the group mean projection, where $\mathbf{J}_c = \mathbf{I}_{n_c} - \frac{1}{n_c}\iota_{n_c}\iota^{'}_{n_c}$. This transformation is conceptually similar to the within-transformation used in fixed-effects panel data models. As $\mathbf{JL} = 0$, the transformed network model is

$$\mathbf{JY} = \phi_1 \mathbf{JGY} + \phi_2 \mathbf{JG}^{\star}\mathbf{Y} + \mathbf{JX}_c\beta_1 + \mathbf{JG}^{\star}\mathbf{X}\beta_2 + \mathbf{JL}\eta + \mathbf{Ju} \tag{4}$$

The econometric network model (4) incorporates the *endogenous effect*, captured by ϕ_1 and ϕ_2, where a scientist's choice may depend on those of her coauthors, the *contextual effect*, captured by the coefficient β_2, where a scientist's outcome may depend on the exogenous characteristics of her coauthors and the effect of idiosyncratic characteristics is captured by β_1. Furthermore, I distinguish between the *aggregate endogenous effect*, captured by the coefficient ϕ_1, and the *average endogenous effect*, captured by the coefficient ϕ_2, as they originate from different economic models with totally different equilibrium implications.

3.2. Estimation strategy

My goal is to consistently estimate the parameters $\theta = (\phi_1, \phi_2, \beta_1^{\intercal}, \beta_2^{\intercal})^{\intercal}$ in Equation (4) considering the potential threats to identification exposed in the following.

3.2.1. Threats to identification

In analyzing the causal estimates of equation (4), I encounter several identification threats. Firstly, feedback loops complicate the analysis, as scientist i influences scientist j and vice versa. Secondly, the potential for non-random sorting of prolific research into communities can lead to biased estimates. Thirdly, coauthorship connexion may be endogenous, formed based on scientists' expected return of collaboration rather than randomly, introducing confounding factors. Lastly, mechanical links between scientists outcomes and community characteristics may result in spurious correlations.

These issues, particularly the endogenous formation of peer groups and mechanical links, relate to what Manski (1993) describes as 'correlated effects.' Correlations in research outputs among coauthors may reflect pre-existing similarities rather than genuine peer effects. Endogenous coauthorship formation is a significant challenge because coauthorships are formed based on strategic choices, resulting in network structures that reflect these choices. This problem can be addressed by modeling the coauthorship formation process using dyadic regression, a method standard in the network formation literature (Graham and De Paula, 2020). Manski (1993) identifies three types of effects influencing behavior: endogenous effects (direct peer influence), contextual effects (influence of peers' characteristics), and correlated effects (common unobserved factors). Using appropriate instruments and fixed effects models can help isolate these effects, mitigate confounding influences, and provide more accurate estimates of peer effects. However introducing fixed effects raises an interpretation issues, if one believe differences in the average output accross communities is due to topic citations specific patterns for instance, should includes fixed effects. However, if the most likely explanation for the difference in terms of productivity accross communities is due to sorting of individuals then fixed effects should not be included.

3.2.2. Identification challenges and instrumental variables

When the adjacency matrix $\mathbf{G}$ is exogenous, linear social interaction models in the from of Equation (4) can be estimated by the 2SLS based on the linear moment condition $\mathbf{Z}^{\intercal}u(\theta) = 0$, where $\mathbf{Z}$ is a matrix of IVs consisting of linearly independent columns of $\mathbf{J}[\mathbf{X}, \mathbf{G}^{\star}\mathbf{X}, \mathbf{G}\mathbf{X}, \mathbf{G}^{\star 2}\mathbf{X}]$, let also $\underline{\mathbf{X}} = \mathbf{J}[\mathbf{G}\mathbf{Y}, \mathbf{G}^{*}\mathbf{Y}, X, \mathbf{G}^{*}\mathbf{X}]$ and

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

The parameters in Equation (4) can be identified , provided that a certain level of intransitivity exists in the network such that the usual rank condition of IV estimators holds (Bramoullé et al., 2009; Liu and Lee, 2010; Liu et al., 2014).

However, if there exists an unobserved individual factor that affects both the output and the coauthorship connexion, then $\mathbf{G}$ is endogenous and the IV matrix $\mathbf{Z}$ is no longer valid.

In this case, the 2SLS method can be remedied by replacing the observed adjacency matrix $\mathbf{G}$ in the IV matrix by a predicted adjacency matrix $\hat{\mathbf{G}}$ based on exogenous covariate (Kelejian and Piras, 2014). In particular, for each pair of individuals i and j in a network, the utility that individual i derives from forming a link with individual j can be expressed as:

$$U_{ij} = \delta_0 + W_{ij}\delta_1 + \epsilon_{ij},$$

where U_{ij} is the utility for scientist i from coauthorship with scientist j, W_{ij} is the deterministic component of the utility, which is a function of observable characteristics of i and j, and ϵ_{ij} is the stochastic component of the utility, capturing unobserved factors that influence the link formation. Assuming ϵ_{ij} is *i.i.d.* type-I extreme value distributed, we then obtain a logistic regression model for the dyadic network formation.

$$\mathbb{P}(g_{ij} = 1) = \frac{\exp(\delta_0 + W_{ij}\delta_1)}{1 + \exp(\delta_0 + W_{ij}\delta_1)}. \tag{5}$$

Based on Equation (5), we define

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

where $\hat{\delta}_0$ and $\hat{\delta}_1$ are convergent estimate. To make sure the predicted adjacency matrix is uniformly bounded in row and column sums, we 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), and define the $(i,j)th$ element of the predicted adjacency matrix $\hat{G}$ as $\hat{g}_{ij}/\hat{d}$ if $i \neq j$ and zero otherwise.

For the "linear in means" part of the econometric model (4), I employ the instrumental variable (IV) strategy outlined by Jochmans (2023). The instrument is constructed by creating a "leave-own-out" subnetwork for each scientist, where all links involving that scientist are removed. This exogenous subnetwork offers predictive insights into the scientist's collaboration behavior, provided certain assumptions are met. Specifically, the self-selection bias should stem from agent-specific unobservables that do not affect the link formation decisions of other pairs of agents. While these unobservables

may exhibit dependence within groups and cause group-level endogeneity, the conditions allow for interdependent link formation as long as such dependencies are limited to within-group interactions. We then use the leave-own-out network to instrument both the average coauthor characteristics and the average coauthor outcomes, based on the corresponding averages within this exogenous network. The construction of instrumental variables for $\mathbf{G}^\star$ 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 defined adjacency matrix $\mathbf{G}-i$, which has been augmented by adding a row and a column of zeros at position i. This adjustment is made for notational clarity and ensures that the resulting matrix retains the same $n \times n$ dimensions as the full network matrix $\mathbf{H}$. The matrix $H-i$ represents the transition matrix for the network with all links involving agent i removed.

For each scientist, I construct the subnetwork by removing all links in which that scientist is involved. This "leave-own-out" network is exogenous and provides valuable predictive information regarding the scientist's own linking behavior. I then instrument the average characteristics of coauthors by using the average of these characteristics in the leave-own-out network. Similarly, I instrument the average coauthor outcomes by using the corresponding averages within this network.

The matrix $\mathbf{G}^\star$ can be interpreted as a transition matrix that gives the probability of moving from scientist i to scientist j in a single step within the network defined by the original adjacency matrix $\mathbf{G}$. The entries of this $n \times n$ 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. $\mathbf{G}^\star\mathbf{X}$ and $\mathbf{G}^\star\mathbf{y}$ can be instrumented by the $\mathbf{Q}_1\mathbf{X}$ and $\mathbf{Q}_2\mathbf{y}$.

The IV matrix based on the predicted adjacency matrix $\hat{\mathbf{G}}$, $\mathbf{Q}_1\mathbf{X}$, $\mathbf{Q}_2\mathbf{y}$ and is denoted by $\hat{Z}$ and includes

linearly independent columns of $\mathbf{J}[\mathbf{X}, \mathbf{Q}_1\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \mathbf{Q}_2\mathbf{X}, \hat{\mathbf{G}}\mathbf{L}]$, the corresponding 2SLS estimator is given by

$$\hat{\theta}_{2sls} = \left[\underline{\mathbf{X}}^{\intercal}\hat{\mathbf{Z}}\left(\hat{\mathbf{Z}}^{\intercal}\hat{\mathbf{Z}}\right)^{-1}\hat{\mathbf{Z}}^{\intercal}\underline{\mathbf{X}}\right]^{-1}\underline{\mathbf{X}}^{\intercal}\hat{\mathbf{Z}}\left(\hat{\mathbf{Z}}^{\intercal}\hat{\mathbf{Z}}\right)^{-1}\hat{\mathbf{Z}}^{\intercal}\mathbf{J}\mathbf{y}. \tag{6}$$

As pointed by Lee et al. (2021), The consistency of the proposed 2SLS estimator does not rely on the consistency of the estimator $\hat{\delta} = \left(\hat{\delta_0}, \hat{\delta_1}^{\intercal}\right)^{\intercal}$.

4. Network and Data

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

4.1. Data Sources

Publication records. My primary data source is the MEDLINE publication database.[1] MEDLINE, the US National Library of Medicine's primary bibliographic database, includes almost all journal articles in life sciences dating back to 1946. It provides a wide range of data elements on journal articles, such as author names, publication dates, journal titles, grant acknowledgments, and Medical Subject Headings (MeSH) terms.

MeSH terms are a controlled vocabulary maintained by the National Library of Medicine, offering a detailed categorization of the intellectual space covered by the biomedical research literature. For our purposes, it is important to note that MeSH terms are assigned to each publication by professional indexers who focus exclusively on the scientific content of the papers.[2] This ensures that MeSH terms provide an objective description of the paper's scientific content.

To obtain citation data for the articles, I use the NIH Open Citation Collection (Hutchins et al., 2019). To focus on the relevant research within oncology, I extracted all articles with at least one relevant MeSH term. Relevant MeSH terms are defined as any child term of "Neoplasms" in the MeSH tree.

[1] We limit the publications to research papers, excluding book reviews, letters, etc.

[2] The Medical Subject Headings Section staff subject specialists are responsible for areas of the health sciences in which they have knowledge and expertise. Specialists from various disciplines are also consulted regarding organizational changes, and close coordination is maintained with specialized vocabularies.

Author name disambiguation. Author names alone are not reliable identifiers for individuals, as multiple authors can share the same name, and authors' names and affiliations may change over time. Author name disambiguation is a longstanding challenge in gathering author-specific publication data, particularly because bibliographic databases typically lack unique individual author identifiers. In the case of MEDLINE, nearly two-thirds of authors face name ambiguity, where their last name and first initial are shared by one or more other authors (Smalheiser et al., 2009). The so-called "John Smith" problem has been the focus of numerous disambiguation algorithms that attempt to identify individuals based on their author names.[3]

Despite efforts to create global author identifiers, such as ORCID and ResearcherID, many articles in MEDLINE, especially those published before 2003 (when the ORCID field was added to PubMed), provide limited author information, typically only the last name, first initial, and, for first authors before 2014, the affiliation.

To address ambiguous author names, I rely on the work of Xu et al. (2020), who developed a PubMed Knowledge Graph (PKG) to tackle the author name disambiguation problem. Their algorithm computes clusters of articles that are likely authored by the same individual. This method has shown outstanding accuracy, especially for NIH-funded scientists. Their approach builds on the probabilistic model developed by Torvik et al. (2005), which assumes that articles written by the same author tend to share more common attributes than articles written by different authors.

Author Attributes. To complement the two primary datasets, I incorporate author-level variables derived from metadata compiled by the Torvik research group and its associated projects. Specifically, I focus on authors' ethnicity, gender, and affiliation.

The *Ethnea* tool is used to determine authors' ethnicity based on their first and last names.[4] Unlike other ethnicity matching algorithms, *Ethnea* primarily reflects an author's nationality, rather than distant ancestral origins, and it also accounts for dual ethnicities that may arise from factors like migration or marriage.

Gender information is assigned through the *Genni* tool, which uses a probabilistic model to match first names to gender. This model also incorporates the ethnicity associated with the last name. For

[3] Author name disambiguation remains one of the major unresolved issues in bibliometrics. Kang et al. (2009) provide an excellent review of the problem and the various approaches developed to address it.

[4] The "Ethnea" tool, developed by the University of Illinois, assigns ethnicity based on name matching. It is available at http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py.

example, the name "Andrea" is identified as female if the associated ethnicity is French, but male if the ethnicity is Italian.

Authors' affiliations are obtained using the *MapAffil* tool, which links PubMed authors' affiliation data to cities and geocodes worldwide. The MapAffil 2016 dataset, based on a snapshot of PubMed from October 2016, connects affiliation strings to specific authors and articles. While PubMed recorded only the first author's affiliation prior to 2014, the MapAffil 2016 dataset extends coverage to include records lacking affiliation information, sourced from databases such as PMC, NIH grants, the Microsoft Academic Graph, and the Astrophysics Data System. It is important to note that the dataset does not include data after 2015 and covers only 62.9% of affiliation instances in PubMed.

In addition to ethnicity, gender, and affiliation, I also record each author's degree(s), the year of degree completion, and whether they have been recognized as prize-winning scientists. This comprehensive set of author attributes enables the construction of detailed profiles, helping to control for demographic, educational, and professional characteristics that influence scientific productivity. Specifically, the inclusion of degree information sheds light on an author's academic background and experience level, while prize recognition highlights individuals who have achieved notable accolades in their field, allowing for the control of highly talented scientists.

4.2. Network Definition

My analysis focuses on authors within the field of cancer research, where collaboration is a common practice and publication output serves as a reliable metric for productivity. Scientific knowledge, particularly in its early stages, is often tacit and confined to small, close-knit groups. As such, collaborations play a crucial role in the exchange of information and the generation of new ideas.

To ensure that my sample reflects actual, active researchers rather than practitioners with occasional or "accidental" publications, I apply several inclusion criteria. Specifically, I restrict the sample to scientists who graduated before 2014 (the earlier of their PhD or MD completion date for those with dual degrees) and who were no longer undergraduate students throughout the study period. Additionally, I require each author to have at least one publication as either the first or last author in the cancer research field. This restriction aligns with a well-established social norm in the life sciences, where first authorship is assigned to the junior author responsible for conducting the research, and last authorship is given to the principal investigator leading the study. The middle authors are credited based on their order in the author list, with recognition typically decreasing as one moves away from either end of the list (Zuckerman, 1968; Nagaoka and Owan, 2014).

In the context of the co-authorship network, two authors are considered linked if they co-authored at least one paper between 2010 and 2014 in the field of cancer. This results in a network with a large central component and several smaller, disconnected components.

To identify tightly-knit groups within the largest connected component of the network, I use a modularity-based community detection algorithm developed by Blondel et al. (2008). This algorithm identifies a partition that approximately maximizes the modularity score, which measures the strength of division of the network into communities.[5] For the purposes of this analysis, I include only those communities with at least 30 authors. I pooled all the communities in the same network, so that each community forms a component of the considered network.

4.3. Variable Definitions

IV Construction:. The method for measuring homophily among co-authors is based on the work of Boschma (2005), who suggested that homophily can be assessed across five dimensions: geographical, cognitive, social, institutional, and cultural. However, due to the high correlation between the geographical and institutional dimensions in our context, I exclude the geographical dimension from our analysis.

- **Cognitive proximity** refers to the shared knowledge and experiences between authors within their field, focusing on their areas of expertise rather than the methodologies or tools used (Nooteboom, 2000). Research overlap, which captures the similarity of research interests, is a key factor influencing collaboration. This is consistent with studies like Fafchamps et al. (2010), which find that co-authorship is more likely when researchers share common research interests. To measure cognitive proximity, I analyzed cancer research articles published in the past five years, linking them to second-level MeSH (Medical Subject Headings) terms. Using these MeSH profiles, I computed cosine similarity (Breschi et al., 2003) between authors' research interests to quantify their cognitive proximity.

- **Social proximity** concerns the relationships between scientists, often forged through previous collaborations (Uzzi, 1996), which reduce uncertainty about potential collaborators' abilities. Breschi and Lissoni (2009) highlights that social ties, such as past co-authorships, play a significant role in facilitating future collaborations. I define social proximity by examining co-author

[5]Modularity-based algorithms are widely studied in fields like computer science and physics (e.g., Loscalzo and Barabási (2016), Ch. 9).

relationships between $t-5$ and $t-1$, considering both direct collaborations and indirect connections (via shared co-authors, with a distance of 2).

- **Institutional proximity** emphasizes the importance of face-to-face interactions and shared institutional resources in fostering scientific collaborations. Studies show that researchers working within the same department or institution are more likely to collaborate (Dahlander and McFarland, 2013; Kabo et al., 2014; Long et al., 2014). To measure institutional proximity, I used MapAffil tools to identify authors' affiliations, designating their primary affiliation during a specific period as the institution where they published the most.

- **Cultural proximity** reflects the influence of shared social norms, language, or cultural backgrounds in promoting collaboration. I used authors' ethnicity as a proxy for cultural proximity, utilizing the Ethnea tool.[6] This aligns with previous studies, such as Dahlander and McFarland (2013) Freeman and Huang (2015), which found that ethnic similarity often leads to higher rates of co-authorship.

In addition to these proximity dimensions, I consider **gender** as an important factor influencing collaboration. Gender can impact co-authorship dynamics in multiple ways. First, the homophily argument suggests that women may prefer to collaborate with other women, and similarly, men with men. Second, gender may also introduce a potential discrimination effect, where male co-authors are preferred over female ones. Finally, gender may influence collaboration decisions due to risk aversion, as discussed by Ductor et al. (2023), this might affect the selection process for entering projects perceived as risky.

Author Productivity. To gather demographic data for each scientist, I rely on CVs, supplemented when necessary by information from professional social networks or faculty websites. This data includes the scientist's degrees (e.g., bachelor's, MD, PhD, or MD/PhD) and the institutions where they obtained these qualifications. Additionally, I use metadata from the Torvik Research Group to determine each scientist's gender.
I also compute the cumulative number of citations each scientist has received up to the start of the study period.

[6]Ethnea maps people into 26 ethnicity groups, including English, Hispanic, Chinese, German, Japanese, French, Italian, Indian, Arab, and others.

To identify early-career prize winners, I consider prestigious scholarships awarded between 1981 and 2010, including the Pew, Searle, Beckman, Rita Allen, and Packard scholarships. These annual awards, given by charitable foundations, provide seed funding to 20 to 40 young academic life scientists each year, representing the highest form of recognition available to early-career researchers in the first two years of their independent careers.

For late-career recognition, I include individuals who were elected as Members of the National Academy of Sciences or the Institute of Medicine between 1970 and 2010. We also account for current and former Howard Hughes Medical Investigators (HHMIs), a distinction given to select mid-career biomedical scientists by the Howard Hughes Medical Institute every three years, based on their potential to make significant contributions to their fields.

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

5. Empirical findings

5.1. First step: collaboration choices

I start my empirical analysis by examining how good is the network formation model I'm using to instrument the observed network. Table (2) presents the estimation results for Model (5), including all explanatory variables from prior literature.

One notable finding is that attending the same graduate institution during the same period does not significantly increase the likelihood of collaboration. This may be due to institutional practices aimed at steering students into distinct research areas, which likely reduces direct competition among alumni entering the job market. On the other hand, the other factors examined in the model yield results

that align with prior expectations.

Regarding social proximity, I find that co-authorship networks are influenced by the previous collaboration network. Specifically, being within one or two steps of a potential collaborator in the previous period network significantly raises the probability of collaboration. This supports the idea that social closeness enhances the likelihood of collaboration, in line with earlier studies that emphasize the role of social proximity. A plausible explanation is that network proximity provides better access to information about the skills and productivity of potential collaborators.

The availability of material resources—particularly those tied to research infrastructure—also appears to drive collaboration, especially within institutions. Researchers at the same institution are more likely to collaborate, likely due to easier access to shared resources like laboratories, hospitals, or advanced technologies (e.g., genome sequencing). Moreover, informal interactions among researchers in close physical proximity may facilitate collaboration, suggesting that future research could further explore the relative contributions of formal versus informal collaboration channels.

Cultural proximity also plays a significant role, with shared ethnicity emerging as an important predictor of collaboration intensity. This finding aligns with the broader literature on homophily in collaboration, suggesting that individuals from similar cultural backgrounds may experience lower communication and coordination costs, making collaboration more efficient. Interestingly, this effect holds true even among highly productive scientist, contrasting with Freeman and Huang (2015)'s findings, where ethnic homophily was absent among leading economists.

Finally, I observe positive assortativity based on seniority within the co-authorship network. This pattern likely arises because senior researchers, having more years of experience, have had more opportunities to collaborate with other senior colleagues.

Finally, it is worth noting that this model exhibits a relatively high McFadden Pseudo R^2, which provides strong evidence that we are unlikely to encounter a weak-IV problem.[7]

5.2. On the root of peer effects

Table (3) reports the estimation results for the general econometric model (3) using alternative estimators that consider whether the adjacency matrix is endogenous or not in their construction. It also reports the first stage F-test statistic and the over-identifying restrictions (OIR) test p-values.

[7]In the case of a weak-IV problem, the GMM approach developed by Liu et al. (2014) can be employed, although some adjustments are necessary to account for heteroscedasticity.

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 (5) 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.

The 2SLS-1 estimates are only valid if the network adjacency matrix $\mathbf{G}$ is exogenous conditional on control variables $\mathbf{X}$ and network fixed effects. I find that the p-values of the OIR test are quite close to the conventional significance level, raising doubts about the exogeneity of $\mathbf{G}$ in the cases of the local-aggregate and general econometric models. In the case of the 2SLS-2 estimator, the p-values provide strong evidence that my IV strategy successfully corrects for the bias arising from the endogeneity of $\mathbf{G}$.

Which mechanisms are driving scientific peer effects? What is the impact of neglecting one significant mechanism on the assessment of the second? The first question can be answered by the statistical significance of ϕ_1 and ϕ_2 in the general network model. My results indicate that scientific peer effects are due to strategic complementarity, even though one might incorrectly conclude that cancer researchers are subject to conformity by estimating a linear-in-means model in isolation.

My findings highlight important caveats in the empirical analysis of scientific peer effects. Peer effects are a complex phenomenon, and their assessment should be approached with caution. If more than

Table 3: Estimation results for the period 2010-2014

	Dependent variable:					
	Weighted number of citations					
	2SLS-1	2SLS-2	2SLS-1	2SLS-2	2SLS-1	2SLS-2
ϕ_1			0.034***	0.043***	0.034***	0.043***
			(0.002105)	(0.003254)	(0.002307)	(0.003329)
ϕ_2	0.831***	0.646**			-0.118	0.218
	(0.217143)	(0.25002)			(0.181637)	(0.235829)
Asinh(Past Production)	0.157***	0.152***	0.11***	0.101***	0.107***	0.106***
	(0.016909)	(0.016301)	(0.013699)	(0.014)	(0.014533)	(0.015423)
Male	0.103	0.101	0.076	0.065	0.076	0.064
	(0.082742)	(0.080077)	(0.067781)	(0.068691)	(0.067724)	(0.070395)
Early Career Prize	-0.399	-0.405	-0.322	-0.362	-0.309	-0.382
	(0.355075)	(0.338583)	(0.277158)	(0.290506)	(0.28513)	(0.287086)
Clinical Concentration	0.665**	0.77**	0.312	0.277	0.316	0.269
	(0.242362)	(0.259928)	(0.203341)	(0.208594)	(0.204357)	(0.210566)
Phd	0.35**	0.398***	0.421***	0.405***	0.435***	0.377***
	(0.11419)	(0.113169)	(0.087583)	(0.091484)	(0.090792)	(0.098084)
MD	0.247*	0.222	0.112	0.06	0.113	0.063
	(0.120047)	(0.11759)	(0.096707)	(0.102343)	(0.096781)	(0.104116)
Decades after graduation	-0.245***	-0.239***	-0.202***	-0.188***	-0.201***	-0.19***
	(0.051252)	(0.049297)	(0.042798)	(0.043236)	(0.043053)	(0.044161)
Late Career Accolade	0.845***	0.827***	0.663***	0.552**	0.669***	0.54**
	(0.17062)	(0.170785)	(0.162825)	(0.176417)	(0.163791)	(0.179158)
Ivy League undergraduate	0.208*	0.225*	0.163*	0.152	0.165*	0.145
	(0.092619)	(0.090728)	(0.079496)	(0.081415)	(0.080136)	(0.082235)
Cragg-Donald Wald F statistic	14.9358	10.2753	14.9358	10.2753	14.9358	10.2753
OIR test p-value	1	1	0.2402	0.9416	0.2567	0.9508
Community fixed-effect	Yes	Yes	Yes	Yes	Yes	Yes
Contextual Variables	Yes	Yes	Yes	Yes	Yes	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 $\hat{\mathbf{G}}$, $\mathbf{Q}_1$, and $\mathbf{Q}_2$ as instruments for the observed adjacency matrix $\mathbf{G}$ and $\bar{\mathbf{G}}$, respectively. Columns 1 and 2 report estimates for the linear-in-means model, columns 3 and 4 for the local-average model, and the last two columns for the mixed model. 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 Undergraduate. 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 $***$, which denote significance at the 10%, 5%, and 1% levels.

one mechanism is driving scientific peer effects, neglecting one of them can produce biased inferential results. I also report the empirical results obtained when separately estimating the local-aggregate and the local-average models. Comparing columns 2–6, it is clear that the local-average peer effect is overstated if the local-aggregate effect is not taken into account, while the "true" peer effect driver remains the same. As evidence by estimates reported in Table (4), these results are consistent accrross period.

Table 4: Estimation results for the period 2006-2010

	Dependent variable:					
	Weighted number of citations					
	2SLS-1	2SLS-2	2SLS-1	2SLS-2	2SLS-1	2SLS-2
ϕ_1			0.039***	0.051***	0.039***	0.053***
			(0.003475)	(0.005569)	(0.003852)	(0.00583)
ϕ_2	2.11***	0.007			0.099	-0.335
	(0.522101)	(0.188452)			(0.365085)	(0.183765)
Asinh(Past Production)	0.192***	0.173***	0.144***	0.139***	0.146***	0.134***
	(0.026678)	(0.01746)	(0.015786)	(0.017092)	(0.016431)	(0.017655)
Male	0.057	0.099	0.098	0.098	0.095	0.101
	(0.144483)	(0.088195)	(0.081457)	(0.084638)	(0.081926)	(0.087496)
Early Career Prize	-0.482	-0.602	-0.7*	-0.677*	-0.692*	-0.678
	(0.346043)	(0.407592)	(0.321116)	(0.332704)	(0.3123)	(0.365641)
Clinical Concentration	1.004**	1.004***	0.612*	0.479	0.615*	0.465
	(0.38222)	(0.276009)	(0.243294)	(0.244834)	(0.24048)	(0.262524)
Phd	0.218	0.185	0.237*	0.245*	0.238*	0.245*
	(0.178986)	(0.116725)	(0.104702)	(0.106874)	(0.104626)	(0.111723)
MD	0.656**	0.237	0.108	0.128	0.131	0.058
	(0.238028)	(0.138166)	(0.116064)	(0.12019)	(0.146388)	(0.131614)
Decades after graduation	-0.29**	-0.302***	-0.255***	-0.265***	-0.256***	-0.265***
	(0.089656)	(0.060916)	(0.054085)	(0.057613)	(0.053831)	(0.059808)
Late Career Accolade	0.201	0.741***	0.586**	0.55**	0.562**	0.63**
	(0.320005)	(0.211292)	(0.183992)	(0.184453)	(0.206351)	(0.204431)
Ivy League undergraduate	-0.076	-0.034	-0.068	-0.081	-0.069	-0.077
	(0.14299)	(0.098023)	(0.090626)	(0.094834)	(0.089664)	(0.100318)
Cragg-Donald Wald F statistic	11.532	6.8611	11.532	6.8611	11.532	6.8611
OIR test p-value	0.9999	1	0.6218	0.9195	0.6896	0.8689
Community fixed-effect	Yes	Yes	Yes	Yes	Yes	Yes
Contextual Variables	Yes	Yes	Yes	Yes	Yes	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 $\hat{\mathbf{G}}$, $\mathbf{Q}_1$, and $\mathbf{Q}_2$ as instruments for the observed adjacency matrix $\mathbf{G}$ and $\bar{\mathbf{G}}$, respectively. Columns 1 and 2 report estimates for the linear-in-means model, columns 3 and 4 for the local-average model, and the last two columns for the mixed model. 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 Undergraduate. 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 $***$, which denote significance at the 10%, 5%, and 1% levels.

6. Concluding remarks

The emphasis in the scientific peer effects literature is that collaboration can increase scientific productivity. Despite numerous well-identified studies on scientific peer effects, the literature has yet to fully explain their origins. A key challenge is our limited understanding of the underlying mechanisms driving the reduced-form estimates.

This paper seeks to address this gap. My focus is coauthorship in cancer research, and we explore the microfoundations of collaboration in the creation of new scientific knowledge. Identifying the nature of scientific peer effects is both essential for policy development and difficult to disentangle empirically. Although several mechanisms have been proposed in the peer effects literature, no previous study has credibly assessed the contributions of both mechanisms.

I employ an instrumental variables (IV) strategy to address the endogeneity of the coauthorship network, enabling us to study peer effects among 'regular' researchers without relying on natural experiments.

My analysis provides a microfoundation for scientific peer effects by exploring and testing different types of utility functions. My results suggest that the linear in means model which was the workhorse of the scientific peer effects was misleading. In this regard, my findings highlight important caveats for empirical analysis of scientific peer effects. These findings not only shed light on the process of knowledge creation but also offer implications for science policy makers on efficient organization of the scientific workforce.

Appendix A. Cancer related papers

To identify cancer-relevant papers, we 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 wide range of malignancies, including solid tumors and hematologic cancers, providing a broad foundation for cancer research. In addition to C04, we 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 en-

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

compassing 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.

Appendix B. Computing Cognitive proximity

To measure the distance, or rather proximity, in intellectual or "ideas space" between pairs of scientists, we 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, our 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, we 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, we 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. We 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.

References

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

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

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

Adhen Benlahlou. Peer effects in cancer research. 2024.

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.

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

Vincent Boucher, Yann Bramoullé, Habiba Djebbari, and Bernard Fortin. Do peers affect student achievement? evidence from canada using group size variation. *Journal of applied econometrics*, 29 (1):91–109, 2014.

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

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.

Antoni Calvó-Armengol, Eleonora Patacchini, and Yves Zenou. Peer effects and social networks in education. *The review of economic studies*, 76(4):1239–1267, 2009.

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.

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.

Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954.

Lee Fleming. Recombinant uncertainty in technological search. *Management science*, 47(1):117–132, 2001.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

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.

Bryan Graham and Áureo De Paula. *The econometric analysis of network data*. Academic Press, 2020.

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.

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

Felichism W Kabo, Natalie Cotton-Nessler, Yongha Hwang, Margaret C Levenstein, and Jason Owen-Smith. Proximity effects on the dynamics and outcomes of scientific collaborations. *Research Policy*, 43(9):1469–1485, 2014.

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.

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.

Michael D König, Claudio J Tessone, and Yves Zenou. Nestedness in networks: A theoretical model and some applications. *Theoretical Economics*, 9(3):695–752, 2014.

Bruno Latour and Steve Woolgar. *Laboratory Life: The Construction of Scientific Facts*. Princeton University Press, 2 edition, 1986.

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.

Xiaodong Liu, Eleonora Patacchini, and Yves Zenou. Endogenous peer effects: local aggregate or local average? *Journal of economic behavior & organization*, 103:39–59, 2014.

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.

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

Alexandre Mas and Enrico Moretti. Peers at work. *American Economic Review*, 99(1):112–145, 2009.

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

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

Bart Nooteboom. Learning by interaction: absorptive capacity, cognitive distance and governance. *Journal of management and governance*, 4:69–92, 2000.

Eleonora Patacchini and Yves Zenou. Juvenile delinquency and conformism. *The Journal of Law, Economics, & Organization*, 28(1):1–31, 2012.

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

Tamotsu Shibutani. Reference groups as perspectives. *American journal of Sociology*, 60(6):562–569, 1955.

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.

Philip Ushchev and Yves Zenou. Social norms in networks. *Journal of Economic Theory*, 185:104969, 2020.

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.

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 Zuckerman. *Scientific Elite: Nobel Laureates in the United States*. Free Press, 1977.

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.