I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11916

Superstar Economists: Coauthorship Networks and Research Output

Chih-Sheng Hsieh
Michael D. König
Xiaodong Liu
Christian Zimmermann

OCTOBER 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11916

Superstar Economists: Coauthorship Networks and Research Output

Chih-Sheng Hsieh
Chinese University of Hong Kong

Michael D. König
CEPR, ETH Zürich and VU Amsterdam

Xiaodong Liu
University of Colorado Boulder

Christian Zimmermann
Federal Reserve Bank of St. Louis and IZA

OCTOBER 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Superstar Economists: Coauthorship Networks and Research Output*

We study the impact of research collaborations in coauthorship networks on research output and how optimal funding can maximize it. Through the links in the collaboration network, researchers create spillovers not only to their direct coauthors but also to researchers indirectly linked to them. We characterize the equilibrium when agents collaborate in multiple and possibly overlapping projects. We bring our model to the data by analyzing the coauthorship network of economists registered in the RePEc Author Service. We rank the authors and research institutions according to their contribution to the aggregate research output and thus provide a novel ranking measure that explicitly takes into account the spillover effect generated in the coauthorship network. Moreover, we analyze funding instruments for individual researchers as well as research institutions and compare them with the economics funding program of the National Science Foundation. Our results indicate that, because current funding schemes do not take into account the availability of coauthorship network data, they are ill-designed to take advantage of the spillover effects generated in scientific knowledge production networks.

JEL Classification: C72, D85, D43, L14, Z13

Keywords: coauthor networks, scientific collaboration, spillovers, key player, research funding, economics of science

Corresponding author:
Christian Zimmermann
Federal Reserve Bank of St-Louis
P.O. Box 442
St. Louis, MO 63166-0442
USA
E-mail: zimmermann@stlouisfed.org

* We thank Pierre-Philippe Combes, Lorenzo Ductor, Bauke Visser, Kenan Huremovic, Marco Van Der Leij, Matt Jackson, David Miller, Joachim Voth, Francis Bloch, Pietro Biroli, Ralph Ossa, Hannes Schwandt, Seth Richards-Shubik, Fabrizio Zilibotti and various seminar participants at the following institutions: University of Zurich, Laval, Amsterdam, and Munich; Jean Monnet University in St-Étienne, Chinese University of Hong Kong, Baptist University of Hong Kong, National Taiwan University, the NSF Conference on Network Science and Economics at Washington University, the workshop on the Economics of Scientific Research at Erasmus University Rotterdam and the Econometric Society Meeting in Barcelona for their helpful comments. Moreover, we thank Adrian Etter and Marc Biedermann for excellent research assistance. Michael D. König acknowledges financial support from the Swiss National Science Foundation through research grant PZ00P1_154957 / 1. The views expressed are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

1. Introduction

Collaborations between researchers in economics have become significantly more important in recent decades. In 1996 multi-authored papers accounted for 50% of all articles published in economics. This number increased to over 75% in 2014 [Kuld and O'Hagan, 2018].[1] Through such collaborations researchers generate spillovers not only to their direct collaboration partners but also indirectly to other researchers who are connected to them within a complex network of collaborations. However, despite the increasing importance of collaborations in the scientific knowledge production process, existing research policies and funding agencies do not take spillovers in the coauthorhsip network into account. The aim of this paper is to develop and structurally estimate a coauthorship network model that allows us to rank researchers and evaluate research funding schemes that take spillovers in the coauthorship network into account.

We build a micro-founded model for scientific knowledge production that generalizes previous ones in the literature by taking complementarities in research efforts between collaborating researchers into account [cf. Ballester et al., 2006; Cabrales et al., 2011; Jackson and Wolinsky, 1996]. We completely characterize the equilibrium outcome when researchers spend effort in multiple and possibly overlapping projects. The equilibrium solution to this model then allows us to rank the impacts of individual researchers on total research output, and design optimal network-based research funding programs.

Based on our economic micro-foundation, we develop an econometric model in which the unobserved effort levels are determined by the Nash equilibrium. The self-selection of researchers into different projects is determined by a matching process that depends on both the researchers' and projects' characteristics [cf. e.g., Chandrasekhar and Jackson, 2012; Friel et al., 2016]. We estimate this model using data for the network of scientific coauthorships between economists registered in the Research Papers in Economics (RePEc) Author Service.[2]

We then propose a novel ranking measure for economists and their departments, which is derived from our economic micro-foundation that explicitly models spillovers between collaborating economists. Our ranking quantifies the endogenous decline in the total research output due to the removal of an economist from the coauthorship network [cf. Ballester et al., 2006; König et al., 2014] and allows us to determine "key players" [cf. Zenou, 2015], or "superstar" economists [cf. Azoulay et al., 2010; Waldinger, 2010, 2012].[3] Taking into account endogenous effort choices of the authors and their interdependencies across the coauthorship network, we

[1] Additional evidence can be found in Ductor [2014].

[2] When two authors claim the same paper in the RePEc digital library, they are coauthors, and the relationship of coauthorship creates an undirected network between them. RePEc assembles the information about publications relevant to economics from over 2000 publishers, including all major commercial publishers and university presses, policy institutions, and pre-prints from academic institutions. See `http://repec.org/` for a general description of the RePEc project.

[3] Note that the effect of hiring superstar scientists on the profitability of firms has been studied in Hess and Rothaermel [2011], Rothaermel and Hess [2007] and Lacetera et al. [2004]. In particular, Rothaermel and Hess [2007] define superstar scientists as researchers who had both published and been cited at a rate of three standard deviations above the mean. In contrast, our measure of superstar scientists takes into account the spillover effects of one scientist on others in a collaboration network.

find that the highest ranked authors are not necessarily the ones with the largest number of citations or that our ranking coincides with other ones used in the literature. This discrepancy is not surprising, as traditional rankings are typically not derived from micro-economic foundations and typically do not take into account the spillover effects generated in scientific knowledge production networks.

Our model further allows us to solve an optimal research funding problem of a planner who wants to maximize total scientific output by introducing research grants into the author's payoff function [Stephan, 1996, 2012]. We study how the funds to different researchers impact aggregate scientific output [cf. König et al., 2014]. We then aggregate researchers by their research institutions and departments, and compute the optimal funding for these institutions [cf. Aghion et al., 2010]. A comparison of our optimal funding policy with the research funding of the economics program of the National Science Foundation (NSF) indicates that there are significant differences, both at the individual and the departmental levels. In particular, we find that our optimal funding policy is significantly positively correlated with the number of coauthors (or degree in the coauthorship network) of an author. In contrast, the NSF awards are not correlated with the degree and they are positively but not significantly correlated with the optimal funding policy. This highlights the importance of the coauthorship network in determining the optimal funding policy, in contrast to the research funding program of the NSF.

There exists a growing literature, both empirical and theoretical, on the formation and consequences of coauthorship networks. On the empirical side, the structural features of scientific collaboration networks have been analyzed in Goyal et al. [2006], Newman [2001a, 2004, 2001b,c,d] and König [2016]. Fafchamps et al. [2010] study predictors for the establishment of scientific collaborations. Ductor et al. [2014] and Ductor [2014] study how these collaborations affect research output of individual authors. At an aggregate level, Bosquet and Combes [2017] estimate the effect of department size on its research output. In contrast to these works we take a structural approach by introducing a production function for the scientific coauthorship network and provide a clear explanation on how co-authorship networks facilitate scientific knowledge production. Moreover, we develop a micro-founded ranking measure of authors and their departments [cf. Azoulay et al., 2010; Liu et al., 2011; Waldinger, 2010, 2012],[4] and investigate optimal research funding policies [cf. De Frajay, 2016; König et al., 2014; Stephan, 2012].

Our paper is further related to the recent theoretical contributions by Baumann [2014] and Salonen [2016], where agents choose time to invest into bilateral relationships. Our model extends the set-ups considered in these papers to allow for investments into multiple projects involving more than two agents. Moreover, in a related paper Bimpikis et al. [2014] analyze firms competing in quantities à la Cournot across different markets with a similar linear-quadratic

[4] There is also a large literature on how to rank authors/departments according to their productivity measured by citations. See for example Perry and Reny [2016], Palacios-Huerta and Volij [2004], Zimmermann [2013] and Lubrano et al. [2003].

payoff specification and allows firms to choose endogenously the quantities sold to each market. In contrast to these authors, the efforts invested by the agents in different projects in our model are strategic complements as opposed to substitutes in their papers.

The paper is organized as follows. Section 2 introduces the scientific knowledge production function and agents' utility function. The policy relevance of our model is illustrated in Section 3; in Section 3.1 we investigate the impact of the removal of an author from the network, while in Section 3.2 we analyze optimal research funding schemes that take into account the spillovers generated across collaborating authors in the network. The empirical implications of the model are discussed in Section 4. The data used for this study are described in Section 4.1, and our econometric methodology is explained in Section 4.2. The matching process of authors and projects is introduced in Section 4.3, a Bayesian estimation method is discussed in Section 4.4 and estimation results are given in Section 4.5. The empirical key player analysis (at both, author and department levels) is then provided in Section 5. Section 6 provides the optimal research funding policy and compares it with the economics funding program by the NSF. Section 7 concludes. The proofs are relegated to Appendix A. More detailed information about the data can be found in Appendix B and some relevant technical material can be found in Appendices C, D, and E. Additional robustness checks are provided in Appendix F.

2. Theoretical Model

2.1. Team Production Function

Let $\mathcal{P} = \{1, \dots, p\}$ denote a set of projects (research papers) and $\mathcal{N} = \{1, \dots, n\}$ denote a set of agents (authors or researchers). The *production function* for project $s \in \mathcal{P}$ is given by[5]

$$Y_s = Y_s(\mathcal{G}) = \sum_{i \in \mathcal{N}} \alpha_i g_{is} e_{is} + \frac{\lambda}{2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N} \setminus \{i\}} f_{ij} g_{is} g_{js} e_{is} e_{js}, \tag{1}$$

where Y_s is the research output of project s, $g_{is} \in \{0, 1\}$ indicates whether agent i participates in project s, e_{is} is the research effort that agent i spends in project s ($e_{is} = 0$ if agent i does not participate in project s), α_i captures the ability of agent i, $f_{ij} \in (0, 1]$ measures knowledge similarity between agents i and j, the spillover-effect parameter $\lambda > 0$ represents complementarity between the research efforts of collaborating agents, and $\mathcal{G}$ stands for the *bipartite* network of authors and projects (cf., Figure 1).

[5] If efforts e_{is} are measured in logarithms, then $Y_s(\mathcal{G})$ corresponds to a *translog production function* [cf. Christensen et al., 1973, 1975]. The translog production function can be viewed as an exact production function, a second-order Taylor approximation to a more general production function, or a second-order approximation to a CES production function; and it has been used, for example, to analyze production in teams [cf. Adams, 2006]. A related specification, without allowing agents to spend effort across different projects, can be found in Ballester et al. [2006] and Cabrales et al. [2011].

2.2. Utility Function

We assume that the *utility function* of agent i is given by

$$U_i = U_i(\mathcal{G}) = \underbrace{\sum_{s\in\mathcal{P}} g_{is}\delta_s Y_s}_{\text{payoff}} - \underbrace{\frac{1}{2}\left(\sum_{s\in\mathcal{P}} g_{is}e_{is}^2 + \phi\sum_{s\in\mathcal{P}}\sum_{t\in\mathcal{P}\backslash\{s\}} g_{is}g_{it}e_{is}e_{it}\right)}_{\text{cost}}, \ \forall i \in \mathcal{N}, \tag{2}$$

where $\delta_s \in (0, 1]$ is a discount factor,[6] and the parameter $\phi > 0$ represents substitutability between the research efforts of the same agent in different projects.[7] This cost is convex if and only if the $p \times p$ matrix $\mathbf{\Phi}$, with diagonal elements equal to one and off-diagonal elements equal to ϕ, is positive definite. The quadratic cost specification includes the convex separable cost specification as a special case when $\phi = 0$. A theoretical model with a similar cost specification but allowing for only two activities is studied in Belhaj and Deroïan [2014] and an empirical analysis is provided in Liu [2014] and Cohen-Cole et al. [2018]. Further, a convex separable cost specification can be found in the model studied in Adams [2006].

The following proposition provides a complete equilibrium characterization of the agents' effort portfolio $\mathbf{e} = [\mathbf{e}_1^\top, \cdots, \mathbf{e}_p^\top]^\top$, with $\mathbf{e}_s = [e_{1s}, \cdots, e_{ns}]^\top$ for $s = 1, \cdots, p$, in the projects they participate. Let

$$\mathbf{W} = \mathbf{G}(\text{diag}_{s=1}^p\{\delta_s\} \otimes \mathbf{F})\mathbf{G}, \qquad \text{and} \qquad \mathbf{M} = \mathbf{G}(\mathbf{J}_p \otimes \mathbf{I}_n)\mathbf{G}, \tag{3}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{G}$ is an np-dimensional diagonal matrix given by $\mathbf{G} = \text{diag}_{s=1}^p\{\text{diag}_{i=1}^n\{g_{is}\}\}$, $\mathbf{F}$ is an $n \times n$ zero-diagonal matrix with the (i, j)-th $(i \neq j)$ element being f_{ij}, and $\mathbf{J}_p$ is an $p \times p$ zero-diagonal matrix with off-diagonal elements equal to one. Let $\rho_{\max}(\mathbf{A})$ denote the spectral radius of a square matrix $\mathbf{A}$.

Proposition 1. *Suppose the production function for each project $s \in \mathcal{P}$ is given by Equation (1) and the utility function for each agent $i \in \mathcal{N}$ is given by Equation (2). Given the bipartite network $\mathcal{G}$, if*

$$|\lambda| < 1/\rho_{\max}(\mathbf{W}) \qquad \textit{and} \qquad |\phi| < 1/\rho_{\max}((\mathbf{I}_{np} - \lambda\mathbf{W})^{-1}\mathbf{M}), \tag{4}$$

then the equilibrium effort portfolio is given by

$$\mathbf{e}^* = (\mathbf{I}_{np} - \mathbf{L}^{\lambda,\phi})^{-1}\mathbf{G}(\boldsymbol{\delta} \otimes \boldsymbol{\alpha}), \tag{5}$$

where $\mathbf{L}^{\lambda,\phi} = \lambda\mathbf{W} - \phi\mathbf{M}$, $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_p]^\top$ and $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_n]^\top$.

[6] If $\delta_s = 1$, then individual payoff from research output Y_s is not discounted. If $\delta_s = 1/\sum_{i\in\mathcal{N}} g_{is}$, then the individual payoff is discounted by the number of agents (coauthors) participating in project s [cf. Hollis, 2001; Jackson and Wolinsky, 1996; Kandel and Lazear, 1992].

[7] For example, Ductor [2014] finds evidence for a congestion externality proxied by the average number of co-authors' papers that has a negative effect on individual academic productivity.

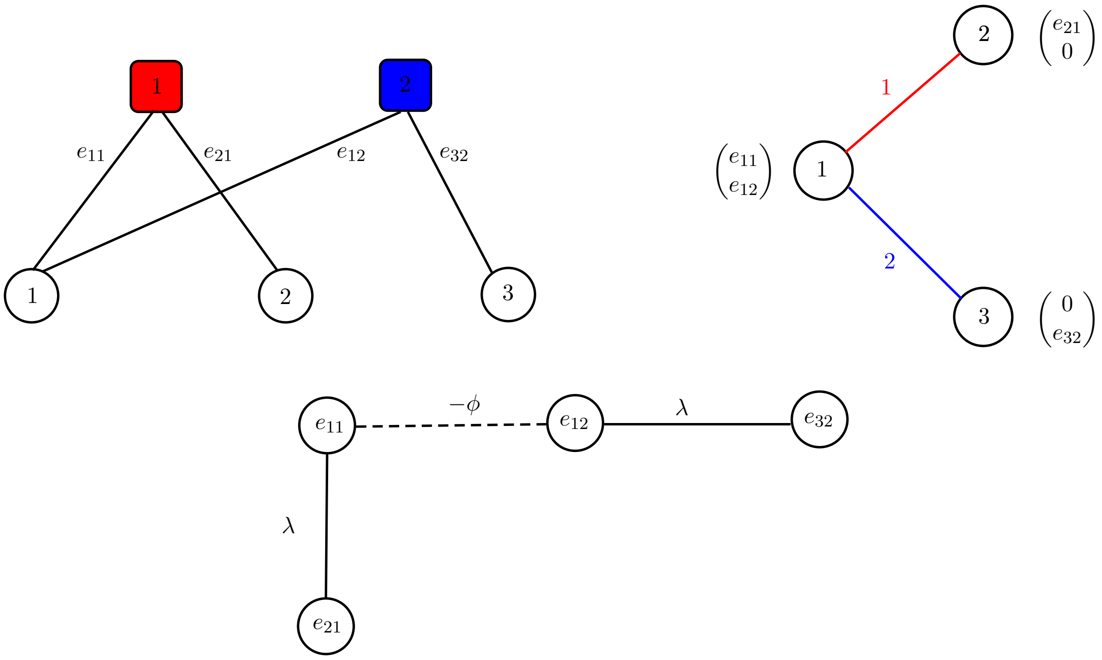

Figure 1: (Top left panel) The bipartite collaboration network $\mathcal{G}$ of authors and projects analyzed in Example 1, where circles represent authors and squares represent projects. (Top right panel) The projection of the bipartite network $\mathcal{G}$ on the set of coauthors. The effort levels of the individual agents for each project they are involved in are indicated next to the nodes. (Bottom panel) The line graph $L(\mathcal{G})$ associated with the collaboration network $\mathcal{G}$, in which each node represents the effort an author invests into different projects. Solid lines indicate nodes sharing a project while dashed lines indicate nodes with the same author.

Observe that the matrix $\mathbf{L}^{\lambda,\phi}$ represents a weighted matrix of the *line graph* $L(\mathcal{G})$ for the bipartite network $\mathcal{G}$,[8] where each link between nodes sharing a project has weight $\lambda\delta_s f_{ij}$, and each link between nodes sharing an author has weight $-\phi$. An example can be found in Figure 1 with $f_{ij} = 1$ for all $i \neq j$ and $\delta_s = 1$ for all s. We will illustrate the equilibrium characterization of Proposition 1 in the following example corresponding to the bipartite network in Figure 1.

Example 1. *Consider a network with* 2 *projects and* 3 *agents, where agents* 1 *and* 2 *are collaborating in the first project and agents* 1 *and* 3 *are collaborating in the second project. An illustration can be found in Figure 1. For expositional purposes, let* $f_{ij} = 1$ *for all* $i \neq j$ *and* $\delta_s = 1$ *for all* s*. Following Equation (3),*

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

[8] Given a network $\mathcal{G}$, its line graph $L(\mathcal{G})$ is a graph such that each node of $L(\mathcal{G})$ represents an edge of $\mathcal{G}$, and two nodes of $L(\mathcal{G})$ are connected if and only if their corresponding edges share a common endpoint in $\mathcal{G}$ [cf. e.g., West, 2001].

and hence

$$\mathbf{L}^{\lambda,\phi} = \lambda\mathbf{W} - \phi\mathbf{M} = \begin{bmatrix} 0 & \lambda & 0 & -\phi & 0 & 0 \\ \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -\phi & 0 & 0 & 0 & 0 & \lambda \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda & 0 & 0 \end{bmatrix}.$$

The nonzero entries of the matrices $\mathbf{W}$ *and* $\mathbf{M}$ *correspond to, respectively, the solid lines and the dashed lines in the line graph depicted in the bottom panel of Figure 1. Thus, the* $(1,2)$*-th and* $(2,1)$*-th elements of the matrix* $\mathbf{L}^{\lambda,\phi}$ *represent the link between* e_{11} *and* e_{21} *with weight* λ *in the line graph, the* $(4,6)$*-th and* $(6,4)$*-th elements represent the link between* e_{12} *and* e_{32} *with weight* λ*, and the* $(1,4)$*-th and* $(4,1)$*-th elements represent the link between* e_{11} *and* e_{12} *with weight* $-\phi$*.*

In this example, the sufficient condition for the existence of a unique equilibrium given by (4) holds if $|\lambda| < 1$ *and* $|\phi| < 1 - \lambda^2$*. From Equation (5) the equilibrium effort portfolio is*

$$\mathbf{e}^* = \begin{bmatrix} e^*_{11} \\ e^*_{21} \\ e^*_{31} \\ e^*_{12} \\ e^*_{22} \\ e^*_{32} \end{bmatrix} = \frac{1}{(1-\lambda^2)^2 - \phi^2} \begin{bmatrix} (1-\lambda^2-\phi)\alpha_1 + \lambda(1-\lambda^2)\alpha_2 - \lambda\phi\alpha_3 \\ \lambda(1-\lambda^2-\phi)\alpha_1 + (1-\lambda^2-\phi^2)\alpha_2 - \lambda^2\phi\alpha_3 \\ 0 \\ (1-\lambda^2-\phi)\alpha_1 - \lambda\phi\alpha_2 + \lambda(1-\lambda^2)\alpha_3 \\ 0 \\ \lambda(1-\lambda^2-\phi)\alpha_1 - \lambda^2\phi\alpha_2 + (1-\lambda^2-\phi^2)\alpha_3 \end{bmatrix}.$$

Observe that

$$\begin{aligned} \frac{\partial e^*_{11}}{\partial \alpha_1} &= \frac{\partial e^*_{12}}{\partial \alpha_1} = \frac{1}{1-\lambda^2+\phi} > 0 \\ \frac{\partial e^*_{21}}{\partial \alpha_1} &= \frac{\partial e^*_{32}}{\partial \alpha_1} = \frac{\lambda}{1-\lambda^2+\phi} > 0 \\ \frac{\partial e^*_{21}}{\partial \alpha_2} &= \frac{\partial e^*_{32}}{\partial \alpha_3} = \frac{1-\lambda^2-\phi^2}{(1-\lambda^2)^2-\phi^2} > 0 \\ \frac{\partial e^*_{11}}{\partial \alpha_2} &= \frac{\partial e^*_{12}}{\partial \alpha_3} = \frac{\lambda(1-\lambda^2)}{(1-\lambda^2)^2-\phi^2} > 0 \end{aligned}$$

which suggest that more-productive agents raise not only their own effort levels but also the effort levels of their collaborators. On the other hand,

$$\begin{aligned} \frac{\partial e^*_{11}}{\partial \alpha_3} &= \frac{\partial e^*_{12}}{\partial \alpha_2} = -\frac{\lambda\phi}{(1-\lambda^2)^2-\phi^2} < 0 \\ \frac{\partial e^*_{21}}{\partial \alpha_3} &= \frac{\partial e^*_{32}}{\partial \alpha_2} = -\frac{\lambda^2\phi}{(1-\lambda^2)^2-\phi^2} < 0 \end{aligned}$$

which suggest that more-productive agents induce lower effort levels spent by agents on other projects. An illustration can be seen in the top panels in Figure 2 with $\alpha_2 = 0.5$*,* $\alpha_3 = 0.8$*,*

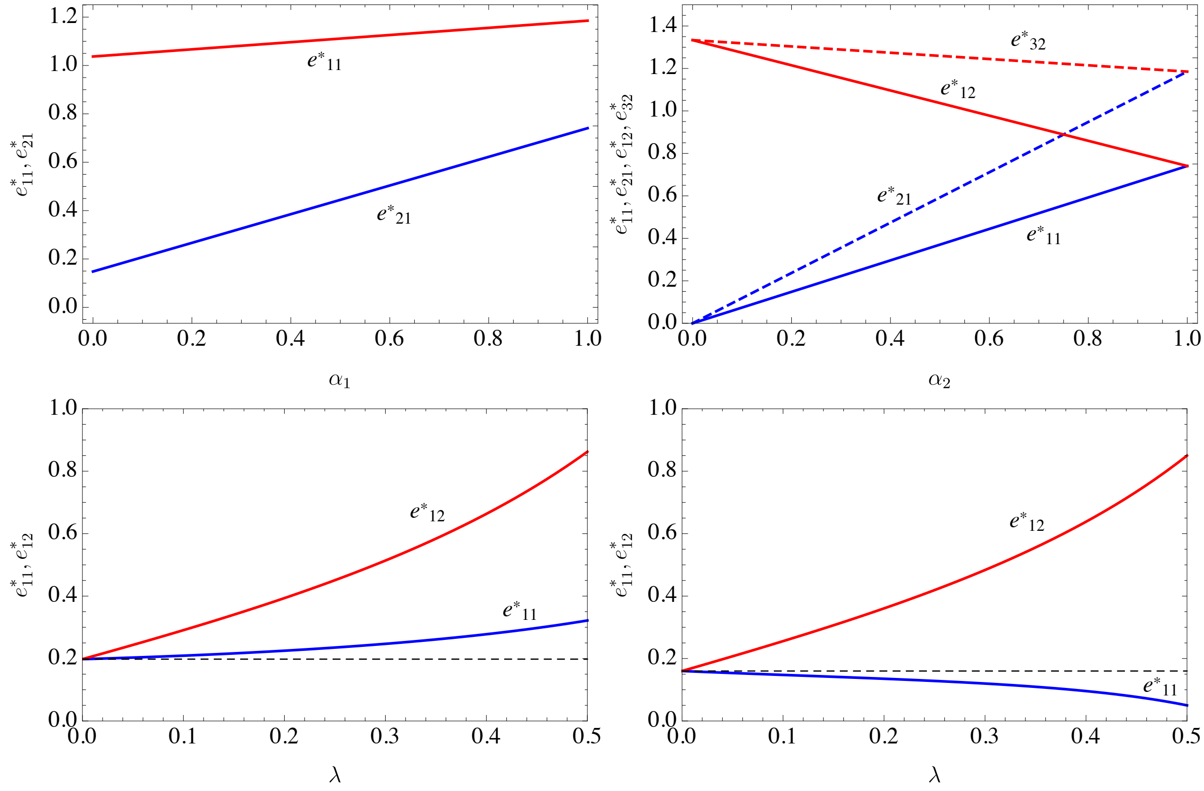

Figure 2: (Top left panel) Equilibrium effort levels for agents 1 and 2 in project 1 for $\phi = 0.75$, $\lambda = 0.25$, $\alpha_2 = \alpha_3 = 1$ (where $e^*_{11} = e^*_{12}$ and $e^*_{21} = e^*_{32}$) and varying values of α_1. (Top right panel) Equilibrium effort levels for agents 1, 2 and 3 in projects 1 and 2 for $\alpha_1 = \alpha_3 = 1$, $\phi = 0.75$, $\lambda = 0.25$ and varying values of α_2. Equilibrium effort levels for agent 1 with $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$, $\phi = 0.05$ (bottom left panel) and $\phi = 0.25$ (bottom right panel) for varying values of λ. The dashed lines in the bottom panels indicate the effort level for $\lambda = 0$.

$\lambda = 0.1$, $\phi = 0.25$ *and varying values of* α_1.

The marginal change of the equilibrium effort e^*_{11} *of agent* 1 *in project* 1 *with respect to the spillover parameter* λ *is given by*

$$\frac{\partial e^*_{11}}{\partial \lambda} = \frac{2\lambda(1-\lambda^2-\phi)^2\alpha_1 + [(1-\lambda^4-\phi^2)(1-\lambda^2)+2\lambda^2\phi^2]\alpha_2 - \phi[(1+3\lambda^2)(1-\lambda^2)-\phi^2]\alpha_3}{[(1-\lambda^2)^2-\phi^2]^2}.$$

Observe that the coefficient of α_3 *is negative. Thus, when* α_3 *is large enough,* $\partial e^*_{11}/\partial\lambda$ *could be negative. The reason is that, with increasing* λ*, the complementarity effects between collaborating agents become stronger, and this effect is more pronounced for the collaboration of agent* 1 *with the more-productive agent* 3*, than with the less-productive agent* 2*. Moreover, when the substitution effect parameter* ϕ *is large, agent* 1 *may spend even less effort in the project with agent* 2*, indicating congestion and substitution effects across projects.*

3. Policy Implications

In the following we analyze the importance of authors and their departments in the coauthorship network (cf. Section 3.1), and we investigate how research funds should optimally be allocated to them (cf. Section 3.2).

3.1. Superstars, Key Players and Rankings

In this section we analyze the impact of a removal of an individual author from the coauthorship network on overall scientific output [cf. e.g., Waldinger, 2010, 2012]. The author whose removal would result in the greatest loss is termed the "key author" [Zenou, 2015] or "superstar" [Azoulay et al., 2010]. More formally, let $\mathcal{G}\backslash\mathcal{A}$ denote the network with agents in the set $\mathcal{A}$ removed from the bipartite network $\mathcal{G}$. The *key author* is defined by[9]

$$i^* \equiv \underset{i\in\mathcal{N}}{\operatorname{argmax}} \left\{ \sum_{s\in\mathcal{P}} Y_s(\mathcal{G}) - \sum_{s\in\mathcal{P}} Y_s(\mathcal{G}\backslash\{i\}) \right\}. \tag{6}$$

Further, aggregating researchers to their departments $\mathcal{D} \subset \mathcal{N}$ allows us to compute the *key department* as

$$\mathcal{D}^* \equiv \underset{\mathcal{D}\subset\mathcal{N}}{\operatorname{argmax}} \left\{ \sum_{s\in\mathcal{P}} Y_s(\mathcal{G}) - \sum_{s\in\mathcal{P}} Y_s(\mathcal{G}\backslash\mathcal{D}) \right\}. \tag{7}$$

3.2. Research Funding

In this section we consider a simple, merit-based research funding policy that takes complementarities in collaborative research efforts into account. For this purpose we consider a two-stage game: in the first stage, the planner announces the research funding scheme that the authors should receive, and in the second stage the authors choose their research efforts, given the research funding scheme. The optimal funding profile can then be found by backward induction.[10] Aggregating the individual funds to the department level also allows us to determine the optimal research funding for departments. For a general discussion of funding of academic research, see Stephan [1996, 2012].

We first solve the second stage of the game. We assume that agent $i \in \mathcal{N}$ receives merit-based research funding, $r_{is} \in \mathbb{R}_+$, per unit of the output she generates in project $s \in \mathcal{P}$. Then the utility function (2) can be extended to

$$U_i(\mathcal{G}) = \sum_{s\in\mathcal{P}} (1 + r_{is}) g_{is} \delta_s Y_s - \frac{1}{2} \left(\sum_{s\in\mathcal{P}} g_{is} e_{is}^2 + \phi \sum_{s\in\mathcal{P}} \sum_{t\in\mathcal{P}\backslash\{s\}} g_{is} g_{it} e_{is} e_{it} \right). \tag{8}$$

The Nash equilibrium effort levels for the utility function in Equation (8) are derived in the following proposition.

Proposition 2. *Let* $\mathbf{R} = \operatorname{diag}_{s=1}^{p}\{\operatorname{diag}_{i=1}^{n}\{r_{is}\}\}$. *Suppose the production function for each project* $s \in \mathcal{P}$ *is given by Equation (1) and the utility function for each agent* $i \in \mathcal{N}$ *is given by*

[9] Note that our model can also be used to measure the potential loss (gain) on research output of a department due to a faculty member leaving (joining) one department for (from) another. This could guide the academic wage bargaining process when professors get an offer from a competing university.

[10] A similar planner's problem in the context of subsidies to R&D collaborating firms has been analyzed in König et al. [2014].

Equation (8). Given the bipartite network $\mathcal{G}$, if

$$|\lambda| < 1/\rho_{\max}((\mathbf{I}_{np}+\mathbf{R})\mathbf{W}) \qquad \textit{and} \qquad |\phi| < 1/\rho_{\max}((\mathbf{I}_{np}-\lambda(\mathbf{I}_{np}+\mathbf{R})\mathbf{W})^{-1}\mathbf{M}), \tag{9}$$

then the equilibrium effort portfolio is given by

$$\mathbf{e}^*(\mathbf{R}) = (\mathbf{I}_{np}-\mathbf{L}^{\lambda,\phi}(\mathbf{R}))^{-1}(\mathbf{I}_{np}+\mathbf{R})\mathbf{G}(\boldsymbol{\delta}\otimes\boldsymbol{\alpha}), \tag{10}$$

where $\mathbf{L}^{\lambda,\phi}(\mathbf{R}) = \lambda(\mathbf{I}_{np}+\mathbf{R})\mathbf{W}-\phi\mathbf{M}$, $\boldsymbol{\delta} = [\delta_1,\cdots,\delta_p]^\top$ and $\boldsymbol{\alpha} = [\alpha_1,\cdots,\alpha_n]^\top$.

Note that compared with Proposition 1, the introduction of research funding $\mathbf{R}$ raises the spillover parameter λ and the abilities $\boldsymbol{\alpha}$ by a factor that is proportional to $(\mathbf{I}_{np}+\mathbf{R})$ in the Nash equilibrium effort levels in Equation (10). Thus, this policy essentially enhances the researchers' abilities and the spillovers generated between collaborators.

Given the equilibrium effort portfolio, $\mathbf{e}^*(\mathbf{R})$, in the first stage of the game, the planner maximizes total output, $\sum_{s\in\mathcal{P}}\sum_{i\in\mathcal{N}} g_{is}\delta_s Y_s(\mathcal{G},\mathbf{R})$, less the cost of the policy, $\sum_{s\in\mathcal{P}}\sum_{i\in\mathcal{N}} r_{is}g_{is}\delta_s Y_s(\mathcal{G},\mathbf{R})$. The planner's problem can thus be written as

$$\begin{aligned} \mathbf{R}^* = & \operatorname*{argmax}_{\mathbf{R}\in\mathbb{R}_+^{np\times np}} \sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{P}}(1-r_{is})g_{is}\delta_s Y_s(\mathcal{G},\mathbf{R}) \\ & \text{s.t. } \mathbf{R} = \operatorname{diag}_{s=1}^{p}\{\operatorname{diag}_{i=1}^{n}\{r_{is}\}\}, r_{is}\in\mathbb{R}_+, \forall i\in\mathcal{N}, \forall s\in\mathcal{P}, \end{aligned} \tag{11}$$

where $Y_s(\mathcal{G},\mathbf{R})$ is the output of project s from Equation (1) with the equilibrium effort levels $\mathbf{e}^*(\mathbf{R})$ given by Equation (10). Equation (11) can then be solved numerically using a constrained nonlinear optimization algorithm [cf. e.g., Nocedal and Wright, 2006].[11]

4. Empirical Implications

4.1. Data

The data used for this study make extensive use of the metadata assembled by the RePEc initiative and its various projects. RePEc assembles information about publications relevant to economics from over 2000 publishers, including all major commercial publishers and university presses, policy institutions, and pre-prints (working papers) from academic institutions. At the time of our data collection, this encompasses 2.6 million records, including 0.82 million

[11] Finding the optimal subsidy program $\mathbf{R}^*$ is equivalent to solving a *bilevel optimization problem* [cf. Bard, 2013], which can be implemented following a two-stage procedure: First, one computes the Nash equilibrium effort levels $\mathbf{e}^*(\mathbf{R})$ that maximize the utilities of Equation (8) as a function of the funding $\mathbf{R}$. Second, one can apply an optimization routine to Equation (11), for example using `MATLAB`'s function `fmincon`. While this would work fine in applications with few agents, it would quickly become inefficient for larger-scale problems. This bilevel optimization problem can be formulated more efficiently as a *mathematical programming problem with equilibrium constraints* (MPEC; see also Luo et al. [1996]), which treats the Nash equilibrium conditions as constraints. This method has recently been proposed to structural estimation problems following the seminal paper by Su and Judd [2012], which further recommends to use the `KNITRO` version of `MATLAB`'s `fmincon` function to improve speed and accuracy.

pre-prints.[12]

In addition, we make use of the data made available by various projects that build on this RePEc data and enhance it in various ways. First, we take the publication profiles of economists registered with the RePEc Author Service (54,000 authors), which include what they have published and where they are affiliated.[13] Second, we extract information about their advisors, students, and alma mater, as recorded in the RePEc Genealogy project.[14] This academic genealogy data has been complemented with some of the data used in Colussi [2017].[15] Third, we use the New Economics Papers (NEP) project to identify which field specific mailing lists through which the papers have been disseminated.[16] NEP has human editors who determine the field in which new working papers belong. We obtain 99 distinct NEP fields. Fourth, we make use of paper download data that is made available by the LogEc project.[17] Fifth, we use citations to the papers and articles as extracted by the CitEc project.[18] Sixth, we use journal impact factors; and author and institution rankings from IDEAS.[19] Finally, we make use of the "Ethnea" tool at the University of Illinois to establish the ethnicity of authors based on the first and last names.[20]

Compared with other data sources, RePEc has the advantage of linking these various datasets in a seamless way that is verified by the respective authors. Author identification is superior to any other dataset as homonyms are disambiguated by the authors themselves as they register and maintain their accounts. While not every author is registered, most are. Indeed, 90% of the top 1000 economists as measured by their publication records for the 1990–2000 period are registered.[21] We believe that the proportion is higher for the younger generation that is more familiar with social networks and online tools and thus more likely to register with online services. Note also that the 54,000 authors on RePEc amount to more than the combined membership of the American Economic Society, the Econometric Society, and the European Economic Association including overlaps (20,152+6,133+3,215=29,500), not all of which may actually be authors.

In terms of publications, RePEc covers all important outlets and over 3,000 journals are listed, most of them with extensive coverage. References are extracted for about 30% of their articles (in addition to working papers) to compute citation counts and impact factors. The missing references principally come from publishers refusing to release them for reasons related to copyright protection. While the resulting gap is unfortunate, it is unlikely to result in a bias against particular authors, fields, or journals. The exception may be authors who are

[12] See http://repec.org/ for a general description of RePEc.

[13] RePEc Author Service: https://authors.repec.org/

[14] RePEc Genealogy project: https://genealogy.repec.org/

[15] We would like to thank Tommaso Colussi for sharing the data with us.

[16] NEP project: https://nep.repec.org/

[17] LogEc project: http://logec.repec.org/

[18] CitEc project: http://citec.repec.org/

[19] IDEAS: https://ideas.repec.org/top/. For a detailed description of the factors and rankings, see Zimmermann [2013].

[20] Ethnea: http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py

[21] https://ideas.repec.org/coupe.html

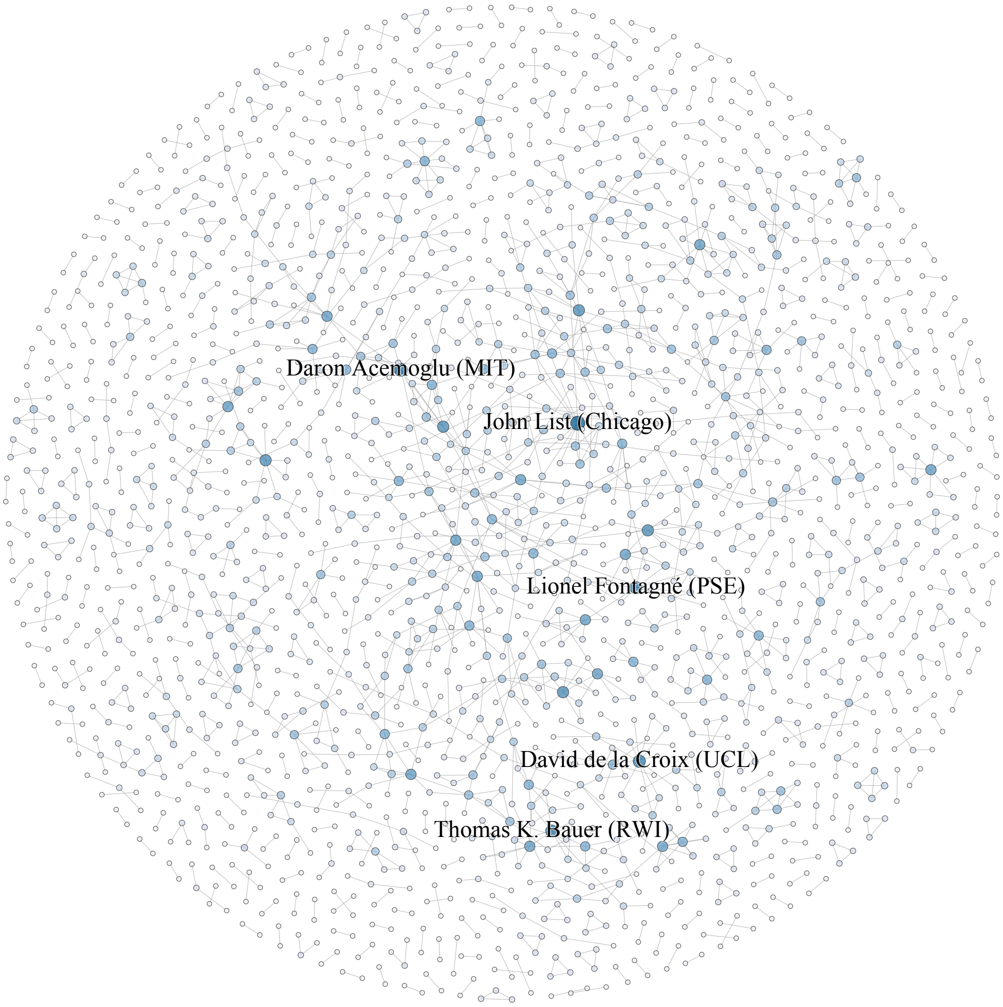

Figure 3: The collaboration network among authors in the RePEc database considering only coauthored projects and dropping projects with zero citations. A node's size and shade indicates its degree. The names of the five authors with the largest number of coauthors (degree) are indicated in the network. These are with decreasing order of degree: John List (University of Chicago), Daron Acemoglu (Massachusetts Institute of Technology; MIT), Thomas K. Bauer (Leibniz Institute for Economic Research; RWI), Lionel Fontagné (Paris School of Economics; PSE), and David de la Croix (Université Catholique de Louvain; UCL).

significantly cited in outlets outside of economics that may or may not be indexed in RePEc (note that several top management, statistics, and political science journals are also indexed).

The amount of RePEc data that is available for this study is overwhelming for the methods we need to adopt to estimate the model. So, we apply a series of filters to reduce the sample size and to obtain records that are complete for our purposes:

1. We select papers that had a first pre-print version within a given span of years. We choose 2010–2012 because it is old enough to give all authors the chance to have added the papers to their profiles and for the papers to have been eventually published in journals. But it is not too old to make sure we have a good-sized sample, as the coverage of RePEc becomes slimmer with older vintages. To examine robustness of our findings with respect to the selection of sample period, we also study the samples of 2007–2009 and 2013–2015 and report the results in Appendix F.
2. We require all authors of the papers to be registered with RePEc.
3. We require that the RePEc Genealogy includes where and under which advisors all authors studied.

4. We require that ethnicity could be determined for all authors.

In the end, we have a dataset for the years 2010 to 2012 with 8,447 papers written by 3,610 distinct authors for which we have complete data. The numbers are similar for the other years.[22] In our empirical model, we use the number of citations of the paper weighted by recursive discounted impact factors of the citing outlet as the measure of a paper's output.[23] Because computing the weighted recursive impact factor requires information about citations, we further drop 2,860 papers that do not have any citations up to July 2018 when retrieving from RePEc. Meanwhile, we also drop 680 authors who only work on these dropped papers without any citations. To understand an author's ability, we use explanatory variables including author's log lifetime citations (at the point of sample collection), decades after receiving Ph.D., dummy variables for being a male, having an NBER affiliation, and graduating from the Ivy League.[24]

The summary statistics of the variables that we use in our empirical model are provided in Table 1. The paper output measure is heavily right-skewed, with the average equal to 6.77 but the maximum equal to 573.23. The average number of authors in each paper is 1.6. The data contain 81% male authors, 5% editors, 10% having an NBER affiliation, and 13.5% Ivy League graduates. The average experience of authors is 1.08 decades after receiving a Ph.D. and the average number of lifetime citations is 221.8. The average number of papers written by an author in the sample period is 3.07. Figure 5 shows the distributions of authors per paper and the number of papers per author, the latter being much more skewed and dispersed over a range of 1 to 74 papers. Moreover, Figure 3 shows the collaboration network among authors and Figure 4 shows the network of collaborations of departments/institutions from the RePEc database. The network of departments is more concentrated among a few central institutions than the network of coauthors. This might stem from the fact that individual authors are constrained in the number of collaborations they can maintain, while these constraints are much less limiting at an aggregate departmental level.

[22] Summary statistics and estimation results covering the years 2007–2009 and 2013–2015 can be found in Appendix F.

[23] The recursive impact factor R_i of journal i is computed as the fixed point of the following system of equations

$$R_i = \frac{\sum_{j \in \mathcal{J}} R_j C_{ij}}{P_i} \frac{\sum_{j \in \mathcal{J}} P_j}{\sum_{j \in \mathcal{J}} R_j P_j}, \ \forall i \in \mathcal{J}, \tag{12}$$

where $\mathcal{J}$ denotes the set of journals, C_{ij} counts the number of citations in journal j to journal i, P_i is the number of all papers/articles in journal i. It is an impact factor where every citation has the weight of the recursive impact factor of the citing journal. All R_i are normalized such that the average paper has an R_i of one. For the discounted recursive impact factor, each citation is further weighted by $1/T$, where T is the age of the citation in years.

[24] A detailed description of these variables can be found in Appendix B.

Table 1: Summary statistics for the 2010-2012 sample.

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive discounted impact factor	0.0001	573.2295	6.7744	17.5537	5587
number of authors (in each paper)	1	5	1.6100	0.7189	5587
Authors					
Log lifetime citations	0	10.6683	5.4018	1.7731	2930
Decades after Ph.D. graduation	-0.6	5.9000	1.0802	1.0487	2930
Male	0	1	0.8116	0.3910	2930
NBER connection	0	1	0.1031	0.3041	2930
Ivy League connection	0	1	0.1352	0.3419	2930
Editor	0	1	0.0509	0.2197	2930
number of papers (for each author)	1	74	3.0700	3.4347	2930

Notes: We drop papers without any citations when extracting from the RePEc database. Authors who only work on these dropped papers are also dropped from the sample.

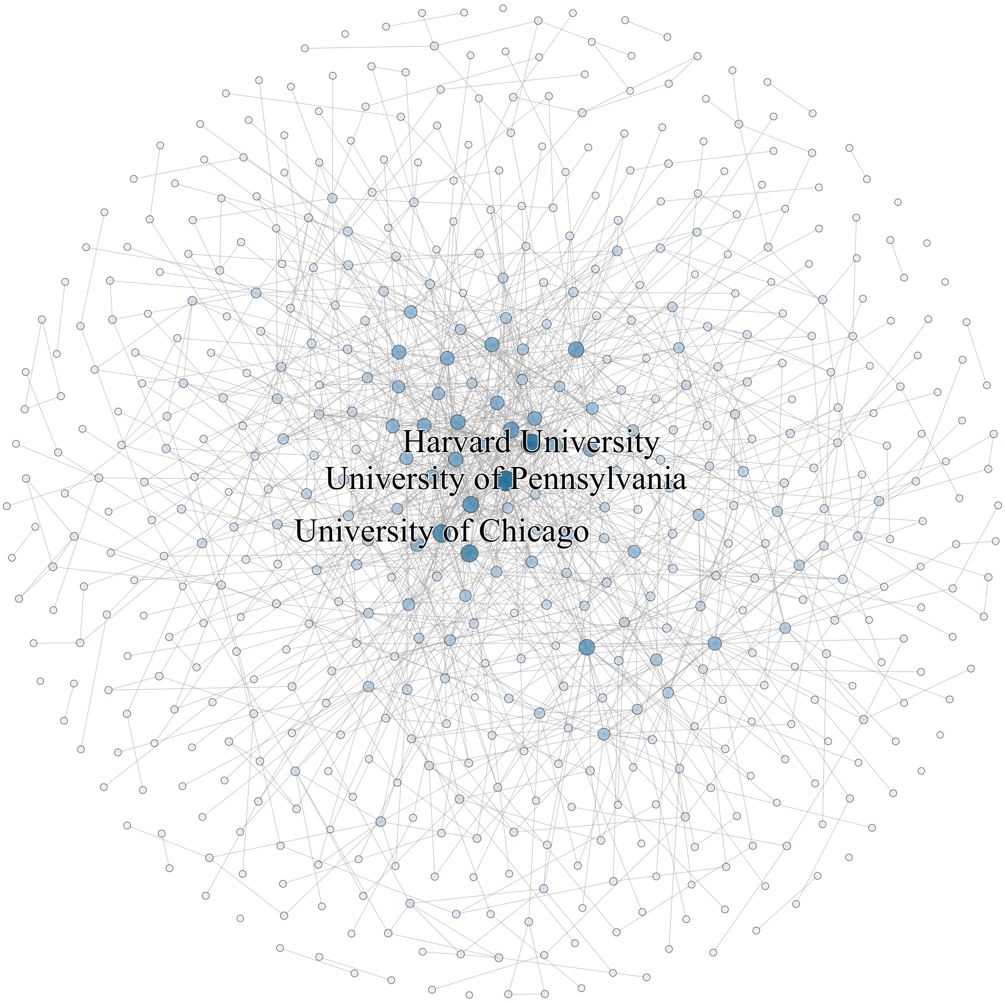

Figure 4: The collaboration network among departments in the RePEc database with a total of 867 unique departments. A node's size and shade indicates its degree. The names of the three departments with the largest degrees are indicated in the network. These are with decreasing order of degree: Harvard University, University of Pennsylvania and the University of Chicago.

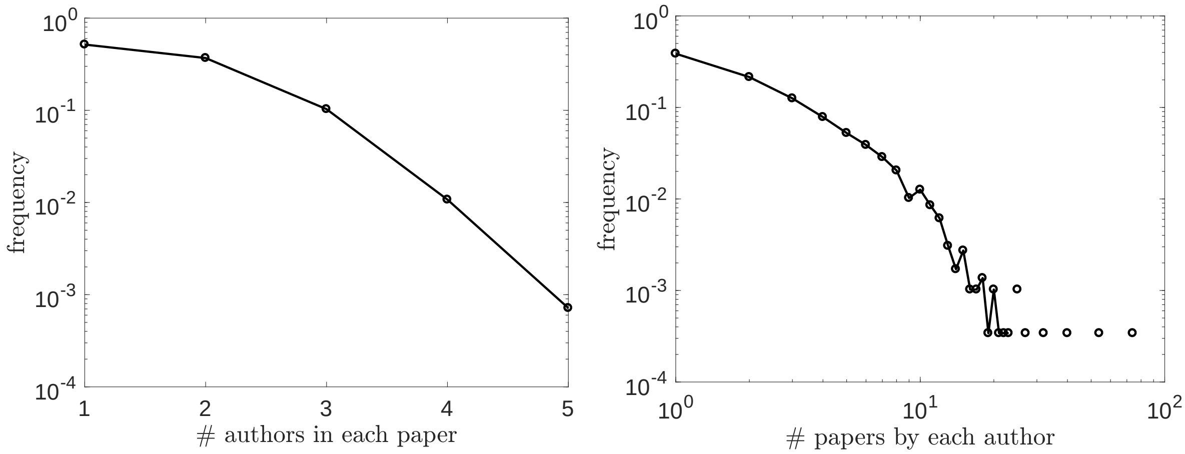

Figure 5: The distribution of authors per paper (left panel) and the number of papers per author (right panel).

4.2. Empirical Production Function

Following Equation (1), the empirical production function of paper $s \in \mathcal{P}$ is given by

$$Y_s = \sum_{i \in \mathcal{N}} \alpha_i g_{is} e_{is} + \frac{\lambda}{2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N} \setminus \{i\}} f_{ij} g_{is} g_{js} e_{is} e_{js} + \epsilon_s, \tag{13}$$

where ϵ_s denotes a paper-specific random shock. We specify $\alpha_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$, where $\mathbf{x}_i$ is a $k \times 1$ vector of author-specific exogenous characteristics, to ensure that author's ability is always presented by a positive value. To estimate this empirical production function, we can consider either the nonlinear least squares method or the likelihood approach (under the distribution assumption of ϵ_s), with the unobervable e_{is} computed by the equilibrium research effort given in Equation (5). However, as the equilibrium research effort portfolio depends on the diagonal matrix $\mathbf{G}$, with its diagonal element $g_{is} \in \{0, 1\}$ indicating whether agent i participates in project s, estimating the empirical production function of Equation (13) may suffer from a *self-selection bias* due to the endogeneity of $\mathbf{G}$. Think of the possibility that high-ability authors may choose high-potential papers to work on. From working on high-potential papers, they also have a better chance to meet other high ability coauthors. As a result, estimating the spillover effect, λ, from Equation (13) without handling endogeneity of the coauthor network $\mathbf{G}$ would suffer from a self-selection bias.

4.3. Matching Process and Identification Strategy

To resolve this self-selection bias, we adopt Heckman's selection-correction approach, or more generally the so-called control function approach [Wooldridge, 2015], in which a selection equation is introduced to model the correlations of error terms between the main output equation and the selection equation. More formally, to address the problem of self-selectivity, we model the endogenous matching process of author $i \in \mathcal{N}$ to paper $s \in \mathcal{P}$ with

$$g_{is} = \mathbb{1}(\psi_{is} + u_{is} > 0), \tag{14}$$

where $\mathbb{1}(\cdot)$ is an indicator function, ψ_{is} captures the matching quality between author i and paper s, and u_{is} is a dyad-specific random component [cf. Chandrasekhar and Jackson, 2012; Friel et al., 2016]. In particular, we assume

$$\psi_{is} = \mathbf{z}_{is}^{\top}\gamma_1 + \gamma_2\mu_i + \gamma_3\kappa_s, \tag{15}$$

where $\mathbf{z}_{is}$ is a $h \times 1$ vector of dyad-specific exogenous variables with its coefficients $\boldsymbol{\gamma}_1$ capturing the similarity between author i and the paper s. We first measure similarity by the research overlap in the NEP fields of paper s and author i. In particular, to avoid the possibility of authors changing fields (and the potential endogeneity concerns that would arise through that), we use the NEP field announced for each author's first paper available in the RePEc database.

In our empirical analysis, we also include additional variables in $\mathbf{z}_{is}$ to capture the average similarity of each author i and the other authors collaborating in project s based on gender, ethnicity, affiliation, whether they have an advisor-advisee relationship [cf. Colussi, 2017], and whether they have coauthored or shared common coauthors in the past.[25] One can interpret the average similarities towards the coauthors' characteristics in the same project as additional way to measure the similarity between the author and the project. For example, if the majority of authors in a project are affiliated to the same department, then this project can be regarded as specific to this department and an author is likely to join this project if she is also affiliated with this department. More generally speaking, our specification tries to reflect the high assortativity in the matching process in scientific coauthorship networks documented in Ductor [2014]. However, note that differently from Ductor [2014], our matching equation allows us to control not only for author but also for paper/project specific effects. The variable μ_i accounts for all author i's time-invariant unobservable attributes, including curiosity, patience, devotion, and others. The variable κ_s similarly represents a paper s's unobservable characteristics. Including μ_i and κ_s allows us to capture the heterogeneity of authors across papers [cf. Graham, 2015, 2017]. Assuming u_{is} is i.i.d. type-I extreme value distributed, we then obtain a logistic regression model for the matching process.

The key feature of the above endogenous matching Equation (14) is to introduce author- and paper-specific latent variables, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_p)$, so that we can also control these latent factors in the determination of paper output. In other words, we extend the production function of Equation (13) to

$$Y_s = \sum_{i\in\mathcal{N}} \underbrace{(\mathbf{x}_i^{\top}\boldsymbol{\beta} + \zeta\mu_i)}_{\alpha_i} g_{is}e_{is} + \frac{\lambda}{2}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}\setminus\{i\}} f_{ij}g_{is}g_{js}e_{is}e_{js} + \underbrace{\eta\kappa_s + v_s}_{\epsilon_s}, \ \forall s \in \mathcal{P}, \tag{16}$$

where v_s is assumed to be independent of u_{is} in Equation (14) and other terms in Equation (16)

[25] We first compare author i with each of her coauthors collaborating in the same project s based on different attributes. The outcome of these comparisons is represented by dummy (indicator) variables with the value one for the same attribute and zero otherwise. We then take the average over the dummy variable as our measure of average similarity.

and normally distributed with zero mean and variance σ_v^2. The identification of the spillover parameter λ in the production function of Equation (16) then comes from the exogenous variation of the research overlap between author i and the potential project s [as in Ductor, 2014]. The research overlap is relevant for the endogenous matching of authors and papers in Equation (14), but is naturally excluded from the production function of Equation (16).

Given $\mathbf{X} = [\mathbf{x}_i]$ and $\mathbf{Z} = [\mathbf{z}_{is}]$, the joint probability function of $\mathbf{Y} = (Y_1, \cdots, Y_p)$ and $\mathbf{G}$ can be specified as

$$\mathbb{P}(\mathbf{Y}, \mathbf{G}|\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}) = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\kappa}} \mathbb{P}(\mathbf{Y}|\mathbf{G}, \mathbf{X}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\kappa}) \mathbb{P}(\mathbf{G}|\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\kappa}) f(\boldsymbol{\mu}) f(\boldsymbol{\kappa}) d\boldsymbol{\mu} d\boldsymbol{\kappa}, \tag{17}$$

from which we can estimate the parameter vector $\boldsymbol{\theta} = (\lambda, \phi, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \eta, \zeta, \sigma_v^2)^\top$, with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \gamma_2, \gamma_3)^\top$. Observe that the author- and project-specific effects, μ_i and κ_s, both appear in the outcome Equation (16) and the matching Equation (15). Thus, omitting them will cause correlations between the error terms of the two equations and hence a self-selection bias emerges. However, by explicitly considering both of them through the joint likelihood of Equation (17), this bias can be corrected for.

4.4. Bayesian Estimation

Since the probability function in Equation (17) involves a high-dimensional integration of latent variables, it is not easy to apply a traditional maximum likelihood method even when resorting to a simulation approach. As an alternative estimation method, the Bayesian Markov Chain Monte Carlo (MCMC) approach can be more efficient for estimating latent variable models [cf. Zeger and Karim, 1991]. We divide the parameter vector $\boldsymbol{\theta}$ and other unknown latent variables into blocks and assign the prior distributions as follows:

$$\begin{aligned}
\mu_i &\sim \mathcal{N}(0, 1), && \text{for } i \in \mathcal{N}, \\
\kappa_s &\sim \mathcal{N}(0, 1), && \text{for } s \in \mathcal{P}, \\
\lambda &\sim \mathcal{N}(0, \sigma_\lambda^2), && \\
\phi &\sim \mathcal{N}(0, \sigma_\phi^2), && \\
\eta &\sim \mathcal{N}(0, \sigma_\eta^2), && \\
\boldsymbol{\xi} &\sim \mathcal{N}_{k+1}(0, \boldsymbol{\xi}_0), && \text{with } \boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \zeta)^\top, \\
\boldsymbol{\gamma} &\sim \mathcal{N}_{h+2}(0, \boldsymbol{\gamma}_0), && \\
\sigma_v^2 &\sim \mathcal{IG}\left(\tfrac{\tau_0}{2}, \tfrac{\nu_0}{2}\right). &&
\end{aligned}$$

We consider the normal and inverse gamma ($\mathcal{IG}$) conjugate priors, which are widely used in the Bayesian literature [Koop et al., 2007]. The hyper parameters are chosen to make the prior distribution relatively flat and cover a wide range of the parameter space, i.e., we set $\sigma_\lambda^2 = \sigma_\phi^2 = \sigma_\eta^2 = 10$, $\boldsymbol{\xi}_0 = 10\mathbf{I}_{k+1}$, $\boldsymbol{\gamma}_0 = 1000\mathbf{I}_{h+2}$, $\tau_0 = 2.2$, and $\nu_0 = 0.1$.

The MCMC sampling procedure combines the Gibbs sampling and the Metropolis-Hastings algorithm. It consists of the following steps:

I. For $i = 1, \ldots, n$, draw the latent variable μ_i using the Metropolis-Hastings algorithm based on $\mathbb{P}(\mu_i|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta}, \boldsymbol{\mu}_{-i}, \boldsymbol{\kappa})$.

II. For $s = 1, \ldots, p$, draw the latent variable κ_s using the Metropolis-Hastings algorithm based on $\mathbb{P}(\kappa_s|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\kappa}_{-s})$.

III. Draw $\boldsymbol{\gamma}$ using the Metropolis-Hastings algorithm based on $\mathbb{P}(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta} \setminus \{\boldsymbol{\gamma}\}, \boldsymbol{\mu}, \boldsymbol{\kappa})$.

IV. Update λ draw the Metropolis-Hastings algorithm based on $\mathbb{P}(\lambda|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta} \setminus \{\boldsymbol{\lambda}\}, \boldsymbol{\mu}, \boldsymbol{\kappa})$.

V. Update ϕ draw the Metropolis-Hastings algorithm based on $\mathbb{P}(\phi|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta} \setminus \{\boldsymbol{\phi}\}, \boldsymbol{\mu}, \boldsymbol{\kappa})$.

VI. Draw $\boldsymbol{\xi}$ using the Metropolis-Hastings algorithm based on $\mathbb{P}(\boldsymbol{\xi}|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta} \setminus \{\boldsymbol{\xi}\}, \boldsymbol{\mu}, \boldsymbol{\kappa})$.

VII. Draw η using the Metropolis-Hastings algorithm based on $\mathbb{P}(\eta|\mathbf{Y}, \mathbf{G}, \boldsymbol{\theta} \setminus \{\boldsymbol{\eta}\}, \boldsymbol{\mu}, \boldsymbol{\kappa})$.

VIII. Draw σ_v^2 using the conjugate inverse gamma conditional posterior distribution.

We collect the draws from iterating the above steps and compute the posterior mean and the posterior standard deviation as our estimation results. In Appendix D we show that the above Bayesian MCMC estimation approach can effectively recover the true parameters from the model of Equations (15) and (16), respectively, in a Monte Carlo experiment.

4.5. Estimation Results

Table 2 reports the estimation results from left to right, considering separately the cases of homogeneous and heterogeneous spillovers (cf. Section 2.1) as well as the case where, in the utility of an author, we discount the payoff by the number of coauthors in each project (cf. Section 2.2). In each case, the first column (Exo. Net.) shows the results where we have assumed that the collaboration network is exogenously given (i.e., ζ and η in Equation (16) are restricted to be zeros), and the estimation procedure is solely based on the production function outlined in Section 4.2. The second column (Endo. Net.) allows the collaboration network to be formed endogenously and is based on the joint estimation of the production function and the matching process described in Section 4.3.

In case of an assumed exogenous network, we find that the spillover effect between coauthors, measured by λ, does not have the expected positive sign. In addition, the congestion (cost) effect across the projects of an author, measured by ϕ, is significant but small in magnitude. In contrast, in the endogenous network case, the estimate of λ is significant and positive (as expected), and the estimate of ϕ is much larger in magnitude. Thus, we can conclude that the estimates of λ and ϕ are downward biased when the endogenous matching between authors and projects is not controlled for. To show why biases are downward, we provide a heuristic explanation in Appendix C by using the estimates from the exogenous and the endogenous network cases to simulate author abilities, efforts, and project outputs. We show that in the exogenous network case author abilities and efforts are overpredicted due to the omission of author- and project-specific effects. This leads to a lower spillover parameter estimate to match

Table 2: Estimation results for the 2010-2012 sample.

		Homogeneous Spillovers		Heterogeneous Spillovers		Discounting # of Coauthors	
		Exo. Net. (1)	Endo. Net. (2)	Exo. Net. (1)	Endo. Net. (2)	Exo. Net. (1)	Endo. Net. (2)
Output							
Spillover	(λ)	-0.0660**	0.0337**	-0.0726	0.0784***	-0.1042*	0.1237**
		(0.0309)	(0.0162)	(0.0511)	(0.0280)	(0.0592)	(0.0553)
Congestion	(ϕ)	0.0202***	0.8656***	0.0187***	0.7847***	0.0203***	0.8791***
		(0.0077)	(0.0645)	(0.0077)	(0.0742)	(0.0081)	(0.0559)
Constant	(β_0)	-0.7237***	-3.2600***	-0.7443***	-3.6207***	-0.7264***	-3.2036***
		(0.1504)	(0.1529)	(0.1485)	(0.1736)	(0.1512)	(0.1431)
Log life-time citat.	(β_1)	0.2653***	0.5901***	0.2649***	0.6250***	0.2655***	0.5705***
		(0.0218)	(0.0195)	(0.0222)	(0.0213)	(0.0222)	(0.0177)
Decades after grad.	(β_2)	-0.1608***	-0.3449***	-0.1582***	-0.3541***	-0.1609***	-0.3609***
		(0.0362)	(0.0209)	(0.0381)	(0.0225)	(0.0372)	(0.0224)
Male	(β_3)	-0.0824	0.3624***	-0.0842	0.4811***	-0.0838	0.4795***
		(0.0724)	(0.0373)	(0.0735)	(0.0425)	(0.0744)	(0.0415)
NBER connection	(β_4)	0.1876***	0.3816***	0.1855***	0.3826***	0.1846***	0.3988***
		(0.0576)	(0.0303)	(0.0571)	(0.0308)	(0.0572)	(0.0310)
Ivy League connect.	(β_5)	0.2288***	0.2751***	0.2323***	0.2294***	0.2319***	0.2093***
		(0.0479)	(0.0313)	(0.0503)	(0.0286)	(0.0489)	(0.0304)
Editor	(β_6)	-0.0985	0.0123	-0.0923	-0.0979*	-0.0955	-0.0242
		(0.0806)	(0.0456)	(0.0790)	(0.0500)	(0.0820)	(0.0443)
Author effect	(ζ)	–	1.7217***	–	1.7814***	–	1.6808***
			(0.0482)		(0.0555)		(0.0434)
Project effect	(η)	–	4.4473***	–	4.1710***	–	4.4002***
			(0.3607)		(0.3425)		(0.3547)
Project variance	(σ_v^2)	261.7498***	132.8909***	262.0227***	134.5656***	261.9254***	131.4169***
		(5.0029)	(2.5674)	(5.0100)	(2.6210)	(5.0022)	(2.5479)
Matching							
Constant	(γ_0)	–	-19.8537***	–	-19.3715***	–	-19.3994***
			(0.2384)		(0.2356)		(0.2345)
Same NEP	(γ_{11})	–	0.4055**	–	0.5081**	–	0.6414***
			(0.2331)		(0.2412)		(0.2374)
Ethnicity	(γ_{12})	–	8.1260***	–	7.7725***	–	7.9508***
			(0.1395)		(0.1300)		(0.1329)
Affiliation	(γ_{13})	–	6.9170***	–	6.6904***	–	6.3189***
			(0.3057)		(0.3067)		(0.3102)
Gender	(γ_{14})	–	4.4114***	–	4.3267***	–	4.2581***
			(0.1356)		(0.1375)		(0.1368)
Advisor-advisee	(γ_{15})	–	9.1362***	–	9.1338***	–	8.8628***
			(0.2257)		(0.2209)		(0.2241)
Past coauthors	(γ_{16})	–	7.9327***	–	7.7202***	–	7.7605***
			(0.1663)		(0.1587)		(0.1577)
Common co-authors	(γ_{17})	–	13.9277***	–	13.5240***	–	13.4888***
			(0.1574)		(0.1534)		(0.1571)
Author effect	(γ_2)	–	2.5689***	–	2.5680***	–	2.5176***
			(0.0582)		(0.0577)		(0.0577)
Project effect	(γ_3)	–	-7.6872***	–	-7.3043***	–	-7.2192***
			(0.1264)		(0.1247)		(0.1212)
Sample size (papers)		5,587		5,587		5,587	
Sample size (authors)		2,930		2,930		2,930	

Notes: The dependent variables include both project output and project-author matching. Model (1) studies project output of Equation (13) assuming exogenous matching between authors and papers. Model (2) studies project output of Equation (16) assuming endogenous matching by Equation (14). We implement MCMC sampling for 30,000 iterations and leave the first 1000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (put into the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

the observed project output levels. In addition, we conduct a Monte Carlo simulation study to investigate the performance of our estimation method. As shown by the simulation results in Appendix D, we find the same downward biases appear in the estimates of λ and ϕ when the collaboration network is incorrectly assumed to be exogenously given.

Regarding the effect of author characteristics on project output, we find that the number of lifetime citations is a positive and significant predictor of research output [cf. e.g., Ductor, 2014], while experience (measured by decades after receiving a Ph.D.) is significantly negative.[26] This finding mirrors Ductor [2014], who shows that career time has a negative impact on productivity and it is consistent with the academics' life-cycle effects documented in Levin and Stephan [1991]. The male dummy shows a positive effect on research output when controlling for network endogeneity [cf. Ductor et al., 2017; Krapf et al., 2017].[27] Being affiliated with the NBER positively and significantly impacts research output. Similarly, having attended an Ivy League university also positively affects output. The editor dummy generally shows insignificant effects on output. The author-specific and project-specific latent variables are found to positively and significantly affect research output. Moreover, the project variance is smaller with the exogenous network case compared with the endogenous case, indicating a better fit of the model to the data (see also Appendix E).

Authors might differ in their competencies and knowledge bases. These differences can affect the spillovers and complementarities authors generate when collaborating on a joint project. In order to capture these heterogeneities, we construct the Jaffe proximity measures of research fields (NEP) between each pair of authors.[28,29] We then incorporate this proximity measure into the production function of Equation (1). In the case of heterogeneous spillovers in Table 2, we again find that when omitting the endogenous matching of authors and papers, the estimate of λ and ϕ are downward biased. However, after controlling for endogenous matching, the estimates of λ and ϕ become significant with the expected signs. Also note that the estimate of λ doubles compared to the homogeneous spillovers case. This is due to the fact that the Jaffe proximity weights are smaller than one and thus a larger spillover coefficient is obtained in compensation.

Similar estimation results can be obtained for the case where we discount the payoff by the number of coauthors in each project in the utility of an author in Equation (2),[30] indicating

[26] Following Rauber and Ursprung [2008] we have also estimated a polynomial of order five in decades after Ph.D. graduation. The result shows that the coefficient of the first order is significantly negative, while the remaining higher orders are insignificant.

[27] In particular, Krapf et al. [2017] find that the effect of parenthood on research productivity is negative for women.

[28] Jaffe [1986] introduces this measure for the analysis of technological proximity between patents. More recently, Bloom et al. [2013] illustrates how Jaffe similarity affects firms' profits with different patent portfolios.

[29] From the authors' NEP fields, we computed their research field proximity following Jaffe [1986] as

$$f_{ij} = \frac{\mathbf{P}_i^\top \mathbf{P}_j}{\sqrt{\mathbf{P}_i^\top \mathbf{P}_i}\sqrt{\mathbf{P}_j^\top \mathbf{P}_j}},$$

where $\mathbf{P}_i$ represents the NEP fields of author i and is a vector whose k-th component P_{ik} counts the number of papers author i has in NEP field k divided by the total number of papers of that author with an attributed field.

[30] However, Kuld and O'Hagan [2018] argue that the available empirical evidence suggests that there is very

the robustness of our results to alternative specifications. Nevertheless, it is still worth pointing out that in the case of discounting utilities, the estimated spillover effect takes a larger value in order to sustain the incentives for authors to collaborate. There are further robustness checks to be found in Section 4.6 below.

For the matching between authors and projects, we indeed find that similarities in the research (NEP) fields positively and significantly affect matchings [Ductor, 2014]. In terms of assortative matching between coauthors, having the same ethnicity, same gender, same affiliation, being past co-authors, and sharing common co-authors all make matching more likely [cf. Freeman and Huang, 2015]. Being in a Ph.D. advisor–advisee relationship also largely contributes to matchings. Further, an author's latent variable shows a significant positive effect on the author–project matching. The project latent variable has a negative effect on the matching, indicating that high-quality projects are more scarce and thus more difficult to join. These results hold across all specifications.

Finally, in Appendix E we examine the goodness-of-fit of the estimated matching model with respect to various network statistics in the data. We find that across the statistics considered, the estimated model is consistent with the observed network.

4.6. Robustness Analyses

We perform a number of robustness checks in Appendix F to gauge the sensitivity of the estimates shown in Table 2. First, in Table F.1 we show the estimation results using only the similarities in the research (NEP) fields for the matching score of Equation (15). The estimated spillover and congestion effects are similar to the ones reported in Table 2, reassuring that identification comes from the exogenous variation in the research overlap between author and projects.

Secondly, the first two columns in Table F.4 show the estimation results with an alternative paper output measure. While in Table 2 we used the sum of citation recursive impact factors to measure a paper's output, in Table F.4 we simply use the sum of the citation impact factors as an alternative. We find that the estimation results are similar to the ones obtained in Table 2. Then, the next four columns of Table F.4 show the estimation results with alternative sample periods, covering the years 2007 to 2009 and the years 2013 to 2015. The corresponding summary statistics for these two sample periods are reported in Tables F.2 and F.3. Similar to the results of Table 2, the estimates of λ and ϕ in the exogenous network case in Table F.4 are downward biased due to omitting the endogenous matching between authors and projects and the biases can be corrected when we jointly model formation of the paper output and the coauthor network. Except the pattern of bias correction, the results further show that the spillover and congestion effects increase over time across sample periods, implying the increasing importance of coauthorship network on economic research.

limited discounting of a published article by the number of co-authors.

5. Rankings for Individuals and Departments

With our estimates from the previous section we are now able to perform various counterfactual studies. In this section we investigate the reduction in total output upon the removal of individual authors or entire research institutions from the network (cf. Section 3.1). Note that when an individual author or an institute is removed, the collaboration network will be rewired after this intervention according to the matching process described in Section 4.3. We use the estimates in Table 2 for the homogeneous spillovers case with endogenous matching for our analysis. The algorithm for network rewiring follows the network simulation method used in the goodness-of-fit examination in Appendix E.

The ranking of individual authors and institutions can be found in Tables 3 and 4, respectively. The key author from our simulation is Robert Barro from Harvard University. Our results suggest that, without this author, total output would be 2.28% lower (cf. Column 9 in Table 3). The second and third highest ranked authors are Ariel Rubinstein from Tel Aviv University and Carmen Reinhart from Harvard University. Their impacts on research output are 1.83% and 1.51%, respectively. In line with the individual ranking, we find that the Department of Economics and the Kennedy School of Government of Harvard University occupy the top two institutions in Table 4.

We find that highly ranked authors tend to have a higher breadth of citing papers across NEP fields (Column 7 in Table 3). Working on a wider range of topics may facilitate the generation of new ideas and start new research projects. A more diverse knowledge base might also help in communicating and collaborating with a broader range of people, and thus allow these authors to occupy more central network positions. Our findings are consistent with Ductor [2014] who shows that a lower degree of specialization has a positive impact on academic productivity. These highly productive authors provide crucial inputs to research projects and cannot easily be substituted in the matching process with their coauthors in the network. Thus, their removal from the network has a strong effect on the total output generated.

Further, from a correlation analysis of the ranking of authors we observe that highly ranked authors tend to have a larger number of projects (Column 2), a larger number of citations (Column 3), and a higher RePEc rank (Column 4). In contrast, purely network-based measures like closeness centrality (Column 5) or betweenness centrality (Column 6) are only weakly correlated with the ranking.[31] The above indicators do not yield the same ranking that we obtain based on our model and the data, as they either are derived from citation counts or depend on the network position only, while our ranking integrates both. Moreover, other rankings are typically not derived from microeconomic foundations and do not take into account spillover effects generated in scientific knowledge production networks.

[31] See the notes in Table 3 for a definition of these measures.

Table 3: Ranking of the top twenty-five researchers from the 2010-2012 sample.

Name	Proj.	Citat.	RePEc Rank[a]	Close.[b]	Betw.[b]	NEP Cites[c]	Organization	Output Loss[d]	Rank
Robert Barro	3	27067	5	4.56	5.22	99.09	Harvard University	-2.28%	1
Ariel Rubinstein	2	4670	301	5.179	1.195	94.04	Tel Aviv University	-1.83%	2
Carmen Reinhart	12	19646	20	4.67	5.47	93.07	Harvard University	-1.51%	3
Oded Galor	11	8132	71	4.864	3.91	93.07	Brown University	-1.21%	4
Nathan Nunn	12	1954	559	4.885	0.537	93.05	Harvard University	-1.11%	5
Sandra Black	5	3149	573	4.76	2.285	95.07	University of Texas-Austin	-1.03%	6
Imran Rasul	11	1735	814	4.58	5.46	90.05	University College London	-0.94%	7
Emmanuel Saez	14	6402	96	4.62	4.62	97.08	University of California-Berkeley	-0.93%	8
Joshua Angrist	6	9553	48	4.55	9.58	98.09	Massachusetts Institute of Technology	-0.92%	9
Michael Waugh	8	518	3000	5.355	0.927	70.03	New York University	-0.91%	10
Lance Lochner	13	2281	930	4.879	2.654	86.05	University of Western Ontario	-0.89%	11
Jorn-Steffen Pischke	9	3717	408	4.69	3.062	98.07	London School of Economics	-0.88%	12
Marc Melitz	9	8111	128	4.77	1.715	96.07	Harvard University	-0.83%	13
Francis Diebold	13	12824	100	4.51	13.55	92.05	University of Pennsylvania	-0.81%	14
Gianmarco Ottaviano	18	5311	234	4.16	38.38	95.07	London School of Economics	-0.80%	15
Michael Keane	11	4675	144	4.45	13.22	95.07	University of New South Wales	-0.80%	16
Justin Wolfers	20	3122	621	4.72	3.54	95.07	University of Michigan	-0.79%	17
Jeffrey Frankel	40	11778	44	4.41	15.33	94.07	Harvard University	-0.76%	18
Paola Giuliano	7	1531	1053	4.64	2.148	90.05	University of California-Los Angeles	-0.75%	19
George Borjas	8	7143	116	4.66	6.70	93.06	Harvard University	-0.75%	20
Susanto Basu	3	2674	706	4.67	2.863	89.05	Boston College	-0.74%	21
Romain Wacziarg	8	3011	602	4.75	1.912	93.06	University of California-Los Angeles	-0.71%	22
Veronica Guerrieri	6	567	2765	4.989	0.418	88.02	University of Chicago	-0.71%	23
Helene Rey	6	2630	636	4.62	3.079	86.04	London Business School	-0.67%	24
Quamrul Ashraf	12	782	2076	5.081	0.603	77.04	Williams College	-0.66%	25

[a] The RePEc ranking is based on an aggregate of rankings by different criteria. See Zimmermann [2013] for more information.

[b] Betweenness centrality measures the fraction of all shortest paths in the network that contain a given node. Nodes with a high betweenness centrality have the potential to disconnect a network if they are removed. In contrast, closeness centrality is a measure of centrality in a network that is calculated as the sum of the length of the shortest paths between the node and all other nodes in the graph. The higher the closeness centrality of a node is, the closer it is to all other nodes in the network. See Wasserman and Faust [1994] and Jackson [2008] for a more detailed discussion of these centrality measures.

[c] NEP cites measures the breadth of citations across NEP fields. Citation breadth is measured by the number k of NEP fields in which at least one paper citing the author has been announced. Ties are broken by computing the number of fields in which x such papers have been announced, where $x = \mod (k/10 + 2)$ (score listed after decimal point).

[d] The output loss for researcher i is computed as $\sum_{s=1}^{p} Y_s(G) - \sum_{s=1}^{p} Y_s(G \backslash i)$ with the parameter estimates from Table 2. See also Equation (3.1) in Section 6.

Table 4: Ranking of the top ten institutions from the 2010-2012 sample.

Organization	Size	RePEc Rank[a]	Output Loss[b]	Rank
Department of Economics, Harvard University	23	1	-6.35%	1
Kennedy School of Government, Harvard University	16	15	-3.72%	2
Economics Department, Brown University	13	17	-2.16%	3
Department of Economics, Northwestern University	13	37	-2.13%	4
Department of Economics, University of California-Berkeley	12	10	-2.03%	5
Economics Department, University of Michigan	17	34	-1.98%	6
Economics Department, Massachusetts Institute of Technology	14	5	-1.93%	7
Department of Economics, Princeton University	12	8	-1.92%	8
Booth School of Business, University of Chicago	11	7	-1.82%	9
Department of Economics, University of Texas-Austin	13	111	-1.73%	10

[a] The RePEc ranking is based on an aggregate of rankings by different criteria. See Zimmermann [2013] for more information.
[b] The output loss for department $\mathcal{D}$ is computed as $\sum_{s=1}^{p} Y_s(G) - \sum_{s=1}^{p} Y_s(G\backslash\mathcal{D})$ with the parameter estimates from Table 2. See also Equation (7) in Section 6.

6. Research Funding for Individuals and Departments

The presence of spillovers in the coauthorship network generates externalities that are not internalized in the utility function of the agents. As a consequence, individual effort levels might not be optimal from a planner's perspective who wants to maximize total research output. In order to create additional incentives for the authors, in this section we analyze a funding scheme that rewards an author in proportion to the output she generates (see Section 3.2).

Assuming that research funds (per unit of output) are homogeneous across projects but heterogeneous across authors, we compute the optimal network-based funding scheme, $(r_i^*)_{1\leq i\leq n}$, by solving Equation (11) with the parameter estimates from Section 4.5.[32] The average optimal funding level per unit of output is 0.0977 with a standard deviation of 0.0732. We find that the optimal research funding policy can raise total net output by 4.2%.

We next compare our optimal funding scheme with funding programs being implemented in the real world [cf. e.g., De Frajay, 2016; Stephan, 2012]. For this purpose we use data on the funding amount, the receiving economics department, and the principal investigators from the Economics Program of the National Science Foundation (NSF) in the U.S. from 1976 to 2016 [cf. Drutman, 2012].[33,34] The economist who had received the largest amount of funds from the NSF is Frank Stafford from the University of Michigan, with total funds amounting to 33 million U.S. dollars. He manages the Panel Study of Income Dynamics (PSID) of U.S. families, which was among the NSF "Top Sixty" overall funded programs in 2010. The average funding amount from the NSF is around $400,000$ U.S. dollars. At the level of organizations and departments, the National Bureau of Economic Research (NBER) received the largest amount of funds totalling to 95 million U.S. dollars,[35] followed by the University of Michigan with a

[32] We use the estimates for the homogeneous spillovers case with endogenous matching in Table 2. Moreover, we initialize the optimization algorithm with the solution from a homogeneous (across authors and projects) funding policy, given by $r^* = 0.0414$. The latter would yield a net output gain of 0.16%.

[33] See https://www.nsf.gov/awardsearch/.

[34] The data coverage before 1976 is incomplete, and we thus discarded years prior to 1976.

[35] The NBER is ranked fourth according to our network-based optimal funding scheme. See also Table 6.

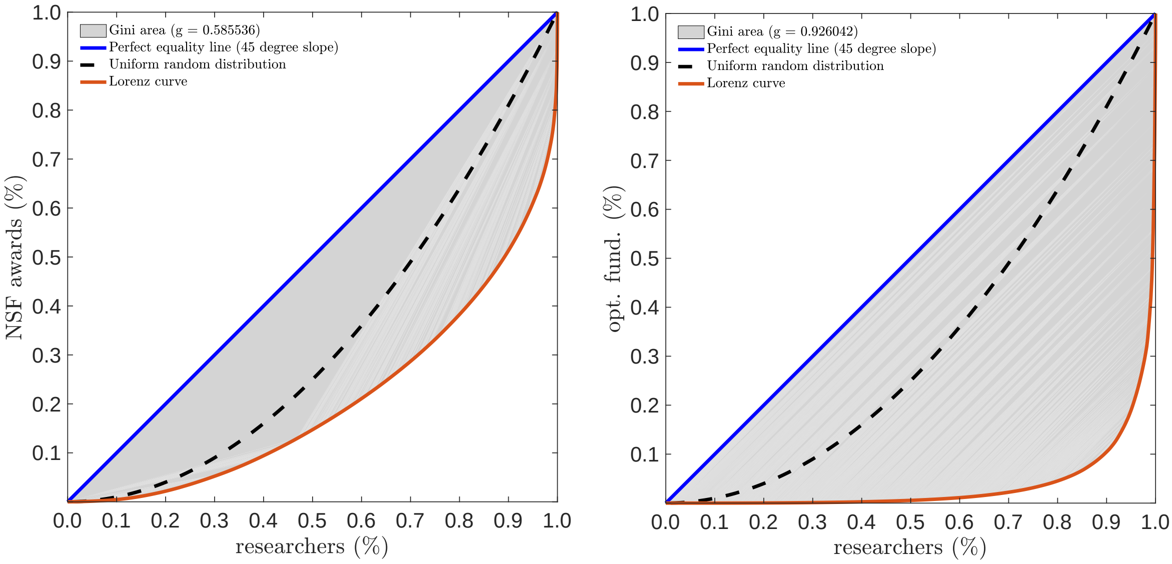

Figure 6: Lorenz curves for the total NSF awards (left panel) and the optimal network-based funding across authors (right panel).

total of 57 million U.S. dollars. The average funding across organizations from the NSF is 2.8 million U.S. dollars. A Lorenz curve illustrating the high inequality of the NSF awards can be seen in the left panel in Figure 6.

The right panel in Figure 6 shows a Lorenz curve of our optimal funding policy. The figure illustrates that the optimal funding policy is highly skewed and tends to concentrate funds towards the most productive authors. The concentration of funds towards the most productive researchers is even higher than for the NSF awards, with a Gini coefficient of $g = 0.59$ for the NSF awards and a coefficient of $g = 0.93$ for our network-based optimal funding policy. The concentration of funds towards the most productive researchers reflects the fact that most of the scientific output is produced by only a small fraction of the most productive economists [Conley and Onder, 2014].

Table 5 shows the optimal network-based research funding amount per author together with the awards these authors actually received from the NSF relative to the total awards provided by the NSF. We observe that the highest ranked/funded authors tend to have a larger number of projects and degree/number of coauthors (Columns 2 and 3 in Table 5; see also Figure 7), illustrating the importance of the coauthorship network for the optimal funding policy. Moreover, the optimal funding amount is negatively correlated with closeness centrality and the RePEc rank and positively correlated with the number of citations and betweenness centrality (Columns 4–7 in Table 5).[36,37] As nodes with a high betweenness centrality tend to disconnect a network when they are removed [cf. e.g., Wasserman and Faust, 1994], the latter indicates that authors bridging different parts of the network should be allocated larger amounts of research funds.

[36] See also Footnote b in Table 3 for a definition and explanation of the closeness and betweenness centrality measures.

[37] Unlike the ranking of authors in Table 5, the optimal funding scheme turns out to be uncorrelated with the breadth of citations across NEP fields as measured by NEP cites.

Table 5: Ranking of the optimal research funding for the top twenty-five researchers for the 2010-2012 sample.[a]

Name	Proj.	Deg.	Citat.	RePEc Rank[b]	Closen.[c]	Between.[c]	NEP Cites[d]	Organization	NSF [%]	Funding [%][e]	Rank
Tim Bollerslev	3	3	17114	60	4.7100	5.7200	57.0100	Duke University	0.1453	7.6192	1
Sendhil Mullainathan	4	2	5704	189	4.6800	1.7315	31.0100	Harvard University	0.0880	4.0190	2
Guido Imbens	1	2	8412	82	4.6000	7.8800	3.0000	Stanford University	0.1490	3.1264	3
David Autor	6	5	6154	151	4.6400	3.5000	89.0500	Massachusetts Institute of Technology	0.2053	3.0803	4
John List	29	11	8911	17	4.1200	111.4500	14.0000	University of Chicago	0.0133	2.7592	5
Daron Acemoglu	17	10	20317	4	4.2200	40.7300	99.0900	Massachusetts Institute of Technology	0.0807	2.0952	6
Shang-Jin Wei	1	2	7374	132	4.3700	19.3300	5.0000	Columbia University	0.0650	1.2921	7
Joshua D Angrist	6	2	9553	48	4.5500	9.5800	0.0000	Massachusetts Institute of Technology	0.2597	1.2440	8
Geert Bekaert	2	2	8440	168	4.8359	4.6800	33.0100	Columbia University	0.0274	1.2235	9
Samuel Kortum	1	2	4633	388	4.8034	2.4685	51.0100	Yale University	0.0678	0.9212	10
Alberto Alesina	13	5	15172	42	4.1800	29.1200	15.0000	Harvard University	0.0031	0.8507	11
Raj Chetty	6	4	3283	224	4.8441	2.5459	2.0000	Harvard University	0.1261	0.7300	12
Nicholas Bloom	13	3	5409	181	4.4500	8.6400	39.0100	Stanford University	0.2982	0.6361	13
David Isaac Laibson	3	2	5390	183	4.4300	11.6900	94.0800	Harvard University	0.0846	0.5314	14
Andrei Shleifer	7	2	42969	1	4.4000	19.7500	95.0800	Harvard University	0.0770	0.5107	15
Pierre Perron	12	6	12462	112	4.7400	11.3200	95.0700	Boston University	0.0885	0.3951	16
Edward Ludwig Glaeser	3	1	12414	47	4.4800	13.2300	49.0100	Harvard University	0.0212	0.3918	17
Parag Pathak	3	3	1504	1120	4.8227	2.4039	40.0100	Massachusetts Institute of Technology	0.2258	0.3885	18
Jonathan Heathcote	6	4	1831	826	5.0537	0.3712	64.0100	Federal Reserve Bank of Minneapolis	0.0451	0.3477	19
Sergio T. Rebelo	11	5	8648	138	4.9358	1.7743	87.0500	Northwestern University	0.0890	0.3353	20
Yuriy Gorodnichenko	8	2	2509	409	4.5500	14.3500	4.0000	University of California-Berkeley	0.0839	0.2308	21
Patrick Kehoe	4	3	7279	129	4.9039	2.6896	18.0000	Stanford University	0.1216	0.1979	22
Frank Schorfheide	11	3	3614	422	4.7300	4.7800	91.0400	University of Pennsylvania	0.1001	0.1640	23
Fabrizio Perri	11	6	2149	754	4.9061	2.3768	67.0200	Federal Reserve Bank of Minneapolis	0.0414	0.1469	24
Xavier Gabaix	8	1	4140	196	4.7800	2.5800	2.0000	Harvard University	0.1378	0.1376	25

[a] We only report the 236 researchers that are listed as principal investigators in the economics program of the National Science Foundation (NSF) in the U.S. from 1976 to 2016 and that can be identified in the RePEc database. The optimal funding policy, however, is computed with data from the full sample.

[b] The RePEc ranking is based on an aggregate of rankings by different criteria. See Zimmermann [2013] for more information.

[c] See also Footnote c in Table 3.

[d] NEP cites measures the breadth of citations across NEP fields. See also Footnote d in Table 3.

[e] The total cost of funds, $\sum_{s=1}^{p} r_{is}^{*} g_{is} \delta_s Y_s(\mathcal{G}, \mathbf{R}^*)$, of researcher i with the optimal research funding scheme $\mathbf{R}^* = (r_{is}^*)$ of Equation (11) in Section 3.2 with the parameter estimates from Table 2 (for the homogeneous spillovers case with endogenous matching).

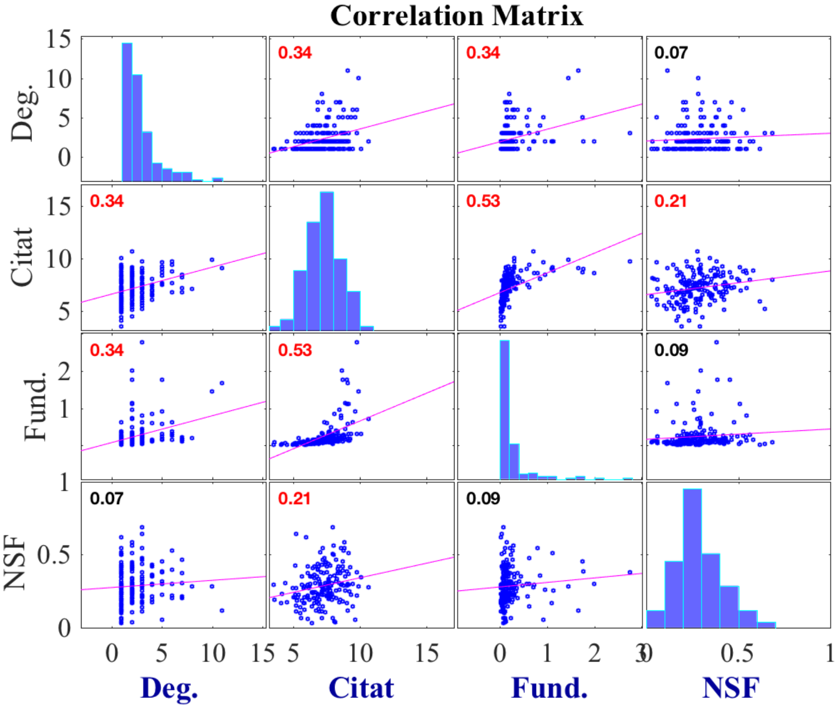

Figure 7: Pair correlation plot of the authors' degrees, citations, total NSF awards, and the optimal funding policy. The Spearman correlation coefficients are shown for each scatter plot, with significant coefficients indicated in bold. The data have been log, respectively, square root transformed to account for the heterogeneity across observations.

Further, we find that the rankings resulted from our optimal funding policy and the rankings chosen by the NSF differ.[38] The author with the highest funds according to our network-based policy is Tim Bollerslev from Duke University (with 7.62% of the total funds), followed by Sendhil Mullainathan from Harvard University (with 4.02% of the total funds) and Guido Imbens from Stanford University (with 3.13% of the total funds). The third-ranked author received almost twice as much funding from the NSF as the second-ranked author. The difference between the optimal network-based funding policy and the one implemented by the NSF is, however, not surprising, as current research funding instruments typically do not take into account the spillover effects generated in scientific knowledge production networks.[39]

The differences between the NSF funding and our funding policy becomes also evident from a simple correlation analysis. Figure 7 shows the correlations of the authors' degrees, lifetime citations, total NSF awards and our network-based optimal funding policy. We observe that the optimal funding policy is significantly positively correlated with the number of citations and the degree (number of coauthors). We also found a positive and significant correlation (0.3) of the key player ranking of Table 3 with the optimal funding policy in Table 5. In contrast, the NSF awards are positively but not significantly correlated with the degree or the optimal

[38] The comparison is based on the 236 authors that could be identified in both the RePEc and the NSF awards databases. The optimal funding policy, however, is computed with data from the full sample.

[39] There are also other systematic differences between our funding policy and the NSF that are worth mentioning. First, the NSF ranking is based on the whole historical record, while our optimal funding ranking is only based on authors' performances between 2010–2012. Second, the NSF funding is application based, while our optimal funding policy is merit (i.e. publication output) based.

Table 6: Ranking of optimal research funding for the top ten departments for the 2010-2012 sample.[a]

Organization	Size	NSF [%]	Funding [%][b]	Rank
Harvard University	49	2.6951	12.8569	1
Duke University	7	2.0840	7.7038	2
Massachusetts Institute of Technology	20	2.1142	7.3352	3
University of Chicago	32	2.7523	6.8529	4
Stanford University	23	4.1304	4.3233	5
Brown University	13	1.1547	3.4792	6
University of California-Berkeley	28	2.1543	3.0588	7
Columbia University	26	2.5537	2.8734	8
Dartmouth College	12	0.5093	1.5955	9
Yale University	21	2.5334	1.3851	10

[a] We only report the 236 researchers that are listed as principal investigators in the Economics Program of the National Science Foundation (NSF) in the U.S. from 1976 to 2016 and that can be identified in the RePEc database. The optimal funding policy, however, is computed with data from the full sample.
[b] The total cost of funds, $\sum_{i\in\mathcal{D}}\sum_{s=1}^{p} r_{is}^{*} g_{is}\delta_s Y_s(\mathcal{G},\mathbf{R}^*)$, for each department $\mathcal{D}$ and researchers $i \in \mathcal{D}$ with the optimal research funding scheme $\mathbf{R}^* = (r_{is}^*)$ of Equation (11) in Section 3.2 with the parameter estimates from Table 2 (for the homogeneous spillovers case with endogenous matching).

funding policy.[40] This highlights the importance of the collaboration network in determining the optimal funding policy, while it does not seem to have an effect on the allocation of NSF awards.

A similar ranking as in Table 5, but at the departmental level, can be found in Table 6. We find that Harvard University receives the largest amount of funds (12.86% of the total), followed by Duke University (7.70% of the total funds). Similar to the ranking of individual authors in Table 5, we observe that the actual funding provided by the NSF does not coincide with the optimal funding policy that we obtain (for example, the fifth-ranked university received twice as much funds from the NSF as the second-ranked one), which explicitly considers spillover effects between the authors within and across different departments.

7. Conclusion

In this paper, we have analyzed the equilibrium efforts of authors who seek to maximize the quality of their scientific output when involved in multiple, possibly overlapping projects with coauthors. We show that, given an allocation of researchers to different projects, the Nash equilibrium can be completely characterized. We then bring our model to the data by analyzing the network of coauthorship between economists registered in the RePEc Author Service. We rank the authors and their departments according to their contribution to aggregate research output, and thus provide a novel ranking measure that is based on microeconomic foundations by determining the key players in the network.

Moreover, we analyze various funding instruments for individual researchers as well as their departments. We show that, because current research funding schemes do not take into account

[40] However, the NSF funding is positively and significantly correlated with the number of citations of an author.

the availability of coauthorship network data, they are ill-designed to take advantage of the spillover effects generated in scientific knowledge production networks. Indeed, the optimal funding policy outcomes deviate substantially from the empirical outcome because we internalize the coauthorship (network) externality. Further, they deviate from the key player rankings because of the different objectives of these two counterfactual studies. While the key player ranking identifies the authors that are already highly productive and exert large spillover effects on their coauthors, the optimal funding policy identifies the authors who can become key players by providing them with additional incentives through a merit based funding scheme.

References

Adams, C. P. (2006). Optimal team incentives with CES production. *Economics Letters*, 92(1):143–148.

Aghion, P., Dewatripont, M., Hoxby, C., Mas-Colell, A., and Sapir, A. (2010). The governance and performance of universities: evidence from Europe and the US. *Economic Policy*, 25(61):7–59.

Azoulay, P., Zivin, J. G., and Wang, J. (2010). Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589.

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

Bard, J. F. (2013). *Practical Bilevel Optimization: Algorithms and Applications.* Berlin: Springer Science.

Baumann, L. (2014). Time allocation in friendship networks. *Available at SSRN 2533533.*

Belhaj, M. and Deroïan, F. (2014). Competing activities in social networks. *The BE Journal of Economic Analysis & Policy*, 14(4):1431-1466.

Bimpikis, K., Ehsani, S., and Ilkilic, R. (2014). Cournot competition in networked markets. *Mimeo, Stanford University.*

Bloom, N., Schankerman, M., and Van Reenen, J. (2013). Identifying technology spillovers and product market rivalry. *Econometrica*, 81(4):1347–1393.

Bosquet, C. and Combes, P.-P. (2017). Do large departments make academics more productive? Agglomeration and peer effects in research. *Journal of Urban Economics*, 101:27–44.

Cabrales, A., Calvó-Armengol, A., and Zenou, Y. (2011). Social interactions and spillovers. *Games and Economic Behavior*, 72(2):339–360.

Chandrasekhar, A. and Jackson, M. (2012). Tractable and consistent random graph models. *Available at SSRN 2150428.*

Christensen, L., Jorgenson, D., and Lau, L. (1973). Transcendental logarithmic production frontiers. *The Review of Economics and Statistics*, 55(1):28–45.

Christensen, L. R., Jorgenson, D. W., and Lau, L. J. (1975). Transcendental logarithmic utility functions. *The American Economic Review*, 65(3):367–383.

Cohen-Cole, E., Liu, X., and Zenou, Y. (2018). Multivariate choices and identification of social interactions. *Journal of Applied Econometrics* , 33(2):165–178.

Conley, J. P. and Onder, A. S. (2014). The research productivity of new PhDs in economics: The surprisingly high non-success of the successful. *Journal of Economic Perspectives*, 28(3):205–216.

Colussi, T. (2017). Social ties in academia: A friend is a treasure. *Review of Economics and Statistics*, 100(1):45–50.

De Frajay, G. (2016). Optimal public funding for research: A theoretical analysis. *RAND Journal of Economics*, 47(3):498–528.

Drutman, L. (2012). How the NSF allocates billions of federal dollars to top universities. *Sunlight foundation blog. https://sunlightfoundation.com/2012/09/13/nsf-funding/.*

Ductor, L. (2014). Does co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics*, 77(3):385–407.

Ductor, L., Fafchamps, M., Goyal, S., and Van der Leij, M. J. (2014). Social networks and research output. *Review of Economics and Statistics*, 96(5):936–948.

Ductor, L., Goyal, S., and Prummer, A. (2017). Gender and social networks. *Working Paper, Middlesex University London.*

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

Freeman, R. B. and Huang, W. (2015). Collaborating with people like me: Ethnic coauthorship within the united states. *Journal of Labor Economics*, 33(S1):289–318.

Friel, N., Rastelli, R., Wyse, J. and Raftery, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences*, 113(24):6629–6634.

Goyal, S., Van der Leij, M. J., and Moraga-Gonzalez, J. L. (2006). Economics: An emerging small world. *Journal of Political Economy*, 114(2):403–412.

Graham, B. S. (2015). Methods of identification in social networks. *Annual Review of Economics*, 7(1):465–485.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063.

Hess, A. M. and Rothaermel, F. T. (2011). When are assets complementary? Star scientists, strategic alliances, and innovation in the pharmaceutical industry. *Strategic Management Journal*, 32(8):895–909.

Hollis, A. (2001). Co-authorship and the output of academic economists. *Labour Economics*, 8(4):503–530.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.

Jackson, M. (2008). *Social and Economic Networks.* Princeton University Press.

Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of Economic Theory*, 71(1):44–74.

Jaffe, A. B. (1986). Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value. *The American Economic Review*, 76(5):pp. 984–1001.

Kandel, E. and Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of political Economy*, 100(4):801–817.

König, M. D. (2016). The formation of networks with local spillovers and limited observability. *Theoretical Economics*, 11:813–863.

König, M. D., Liu, X., and Zenou, Y. (2014). R&D networks: Theory, empirics and policy implications. Forthcoming in the Review of Economics and Statistics.

Koop, G., Poirier, D. J., and Tobias, J. L. (2007). *Bayesian econometric methods.* Cambridge University Press.

Krapf, M., Ursprung, H. W., and Zimmermann, C. (2017). Parenthood and productivity of highly skilled labor: Evidence from the groves of academe. *Journal of Economic Behavior & Organization*, 140:147–175.

Kuld, L. and O'Hagan, J. (2018). Rise of multi-authored papers in economics: Demise of the "lone star" and why? *Scientometrics*, 114(3):1207–1225.

Lacetera, N., Cockburn, I. M., and Henderson, R. (2004). Do firms change capabilities by hiring new people? A study of the adoption of science-based drug discovery. in Joel A.C. Baum, Anita M. McGahan (ed.) Business Strategy over the Industry Lifecycle. *Advances in Strategic Management*, 21:133–159.

Levin, S. G. and Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. *The American Economic Review*, 81(1):114–132.

Liu, X. (2014). Identification and efficient estimation of simultaneous equations network models. *Journal of Business & Economic Statistics*, 32(4):516–536.

Liu, X., Patacchini, E., Zenou, Y., and Lee, L. (2011). Criminal networks: Who is the key player? *Research Papers in Economics, Stockholm University.*

Lubrano, M., Bauwens, L., Kirman, A., and Protopopescu, C. (2003). Ranking economics departments in Europe: a statistical approach. *Journal of the European Economic Association*, 1(6):1367–1401.

Luo, Z.-Q., Pang, J.-S., Ralph, D., 1996. *Mathematical programs with equilibrium constraints.* Cambridge University Press.

Newman, M. (2001a). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409.

Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(90001):5200–5205.

Newman, M. E. J. (2001b). Scientific collaboration networks i. Network construction and fundamental results. *Physical Review E*, 64(1):016131.

Newman, M. E. J. (2001c). Scientific collaboration networks. ii. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132.

Newman, M. E. J. (2004). Who is the best connected scientist? A study of scientific coauthorship networks. *Complex networks*, Springer, Berlin, Heidelberg, 337–370.

Nocedal, J. and Wright, S. (2006). *Numerical optimization.* Springer.

Palacios-Huerta, I. and Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, 72(3):963–977.

Perry, M. and Reny, P. J. (2016). How to count citations if you must. *The American Economic Review*, 106(9):2722–2741.

Rauber, M. and Ursprung, H. W. (2008). Life cycle and cohort productivity in economic research: The case of germany. *German Economic Review*, 9(4):431–456.

Rothaermel, F. T. and Hess, A. M. (2007). Building dynamic capabilities: Innovation driven by individual-, firm-, and network-level effects. *Organization Science*, 18(6):898–921.

Salonen, H. (2016). Equilibria and centrality in link formation games. *International Journal of Game Theory*, 45(4):1133–1151.

Snijders, T. A. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Stephan, P. E. (1996). The economics of science. *Journal of Economic literature*, 34(3):1199–1235.

Stephan, P. E. (2012). *How economics shapes science.* Harvard University Press.

Su, C.-L., Judd, K. L., 2012. Constrained optimization approaches to estimation of structural models. *Econometrica* 80(5): 2213–2230.

Waldinger, F. (2010). Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany. *Journal of Political Economy*, 118(4):787–831.

Waldinger, F. (2012). Peer effects in science: evidence from the dismissal of scientists in Nazi Germany. *The Review of Economic Studies*, 79(2):838–861.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* Cambridge University Press.

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

West, D. B. (2001). *Introduction to Graph Theory.* Prentice-Hall.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86.

Zenou, Y. (2015). Key players. *Oxford Handbook on the Economics of Networks.* Y. Bramoulle, B. Rogers and A. Galeotti (Eds.), Oxford University Press.

Zimmermann, C. (2013). Academic rankings with RePEc. *Econometrics*, 1(3):249–280.

Appendix

A. Proofs

Proof of Propositions 1 and 2. First, we prove Proposition 2. Substitution of Equation (1) into Equation (8) gives

$$
\begin{aligned}
U_i = \sum_{s \in \mathcal{P}} (1 + r_{is}) g_{is} \delta_s \left(\sum_{j \in \mathcal{N}} \alpha_j g_{js} e_{js} + \frac{\lambda}{2} \sum_{j \in \mathcal{N}} \sum_{k \in \mathcal{N} \setminus \{j\}} f_{jk} g_{js} g_{ks} e_{js} e_{ks} \right) \\
- \frac{1}{2} \left(\sum_{s \in \mathcal{P}} g_{is} e_{is}^2 + \phi \sum_{s \in \mathcal{P}} \sum_{t \in \mathcal{P} \setminus \{s\}} g_{is} g_{it} e_{is} e_{it} \right) . \qquad (18)
\end{aligned}
$$

Observe that $e_{is} = g_{is} e_{is}$, as $e_{is} = 0$ if agent i does not participate in project s. The first-order condition of maximizing utility in Equation (18) with respect to $\widetilde{e}_{is}$ gives

$$
e_{is} = (1 + r_{is}) g_{is} \left(\delta_s \alpha_i + \lambda \delta_s \sum_{j \in \mathcal{N} \setminus \{i\}} f_{ij} g_{js} e_{js} \right) - \phi g_{is} \sum_{t \in \mathcal{P} \setminus \{s\}} g_{it} e_{it} .
$$

In matrix form, the first-order condition can be written as

$$
\mathbf{e} = (\mathbf{I}_{np} + \mathbf{R}) \mathbf{G} (\boldsymbol{\delta} \otimes \boldsymbol{\alpha}) + \lambda (\mathbf{I}_{np} + \mathbf{R}) \mathbf{W} \mathbf{e} - \phi \mathbf{M} \mathbf{e}.
$$

If $|\lambda| < 1/\rho_{\max}((\mathbf{I}_{np} + \mathbf{R})\mathbf{W})$, then the matrix $(\mathbf{I}_{np} - \lambda(\mathbf{I}_{np} + \mathbf{R})\mathbf{W})$ is nonsingular. If, in addition, $|\phi| < 1/\rho_{\max}((\mathbf{I}_{np} - \lambda(\mathbf{I}_{np} + \mathbf{R})\mathbf{W})^{-1}\mathbf{M})$, then the matrix $(\mathbf{I}_{np} - \mathbf{L}_r^{\lambda,\phi})$ is nonsingular. Thus, the equilibrium effort levels are given by Equation (10). The proof of Proposition 1 follows the same argument with $r_{is} = 0$. ∎

B. Data Appendix

We use the following variables, retrieved in July 2018:

- Individual author characteristics

  1. Number of lifetime citations to all their works in their RePEc profile.
  2. Number of times their works have been downloaded in the past 12 months from the RePEc services that report such statistics on LogEc (EconPapers, IDEAS, NEP, and Socionet).
  3. Current RePEc ranking of the author. We use the aggregate ranking for the lifetime work.[41]
  4. Current RePEc ranking for the main affiliation of the author.

[41] See https://ideas.repec.org/top/top.person.all.html for the top-ranked economists.

5. Year of completion of terminal degree, as listed in the RePEc Genealogy.
6. Number of registered coauthors during career.
7. Dummy for editor of journal.
8. Dummy for NBER or CEPR affiliation.
9. Dummy for terminal degree from an Ivy League institution.
10. Dummy for main affiliation in the United States.
11. Gender as determined by a likelihood table using the first and possibly middle name. Uncertain matches were almost all resolved through internet search.
12. Ethnicity.
13. Closeness centrality measure.
14. Betweenness centrality measure.
15. Number of NEP fields in which author's work has been cited, to measure breadth of citations.
16. Fields of work, as determined by the NEP fields for which their working papers were selected for email dissemination.
17. First NEP field recorded in career.

- Potential author pair characteristics

1. Co-authorship previous to the period under consideration.
2. Student-advisor relationship, as recorded in the RePEc Genealogy.
3. Joint alma mater of terminal decree as recorded in the RePEc Genealogy.
4. Joint affiliation, taken from the affiliations authors recorded in the RePEc Author Service. As authors may have multiple affiliations, we use two versions: one with only the main affiliation matching for the author-pair, the other where any of the affiliation matches.
5. Joint ethnicity.
6. Joint country of main affiliation.
7. Joint field of work. There are two ways we determine this, both based on the NEP fileds in which the authors published. For the first, we only consider the fields in which each author has written at least four papers or, for authors with less than 10 years of experience, a quarter of all papers announced through NEP. A match is called if at least one field coincides in the author pair. For the second, we consider for each author the proportion of papers in each fields and then compute a score by multiplying the vectors of the authors across all fields.

- Paper characteristics

1. Number of citations for all versions of the paper.
2. Same, but weighted simple impact factors, as listed on IDEAS.

3. Same, but weighted recursive impact factors, as listed on IDEAS.[42]
4. Same, but weighted discounted impact factors, as listed on IDEAS.
5. Same, but weighted recursive discounted impact factors, as listed on IDEAS.
6. Same, but weighted simple discounted impact factors, as listed on IDEAS.
7. If published, the journal's simple impact factor, as listed on IDEAS.
8. If published, the journal's recursive impact factor, as listed on IDEAS.
9. If published, the journal's H-index, as listed on IDEAS.
10. The number of downloads in the past 12 months, as provided by LogEc.
11. The number of authors.
12. Year of publication in a journal.

C. Heuristic Explanation of the Estimation Bias

In this section we provide an explanation why in Table 2 the estimates of λ and ϕ in the case of the exogenous network of Model (1) are biased downward compared with the results in the case of the endogenous network of Model (2). A formal derivation of these estimation biases would be difficult due to the nonlinearity of equilibrium research efforts in Equation (5) used during estimation. As an alternative, we use a counterfactual simulation to study the direction of the bias.

To begin our investigation, first note that the difference between Model (1) and the output equation of Model (2) lies in the presence of author- and project-specific effects in Model (2). In Model (1), where the coauthorship network is regarded as exogenous, we cannot control author- and project-specific effects in the output equation because, for each project and each author who only participated in one project, they only appear once in the sample, and thus we do not have enough variation to identify author- and project-specific effects. On the contrary, Model (2) exploits not only variations across project outputs but also the endogenous matching between authors and projects. The multiple matching outcomes for each author and each project provide sufficient sample variations to identify author- and project-specific effects in Model (2). That is to say, the potential bias problem in the coefficient estimates in Model (1) could be attributed to the omission of author- and project-specific effects.

We first look at the distributions of author ability ($\boldsymbol{\alpha}$), research effort ($\mathbf{e}$), and estimated project output based on the estimates of Model (1) and the estimates of Model (2), respectively. The distributions of estimated author abilities in the 2010-2012 sample from Model (1) and Model (2) are shown in Figure C.1. One can see that the estimated author abilities from Model (2) are generally lower than those from Model (1), indicating that ignoring author-specific effects ($\boldsymbol{\mu}$) results in an upward bias on the estimated author abilities in Model (1). Next, we show the distributions of research efforts from Model (1) and Model (2) in Figure C.2. Not surprisingly, the estimated research effort displays the same pattern as the author ability in Figure C.1,

[42] For the details of the computation of the weighted recursive impact factor, see Footnote 23.

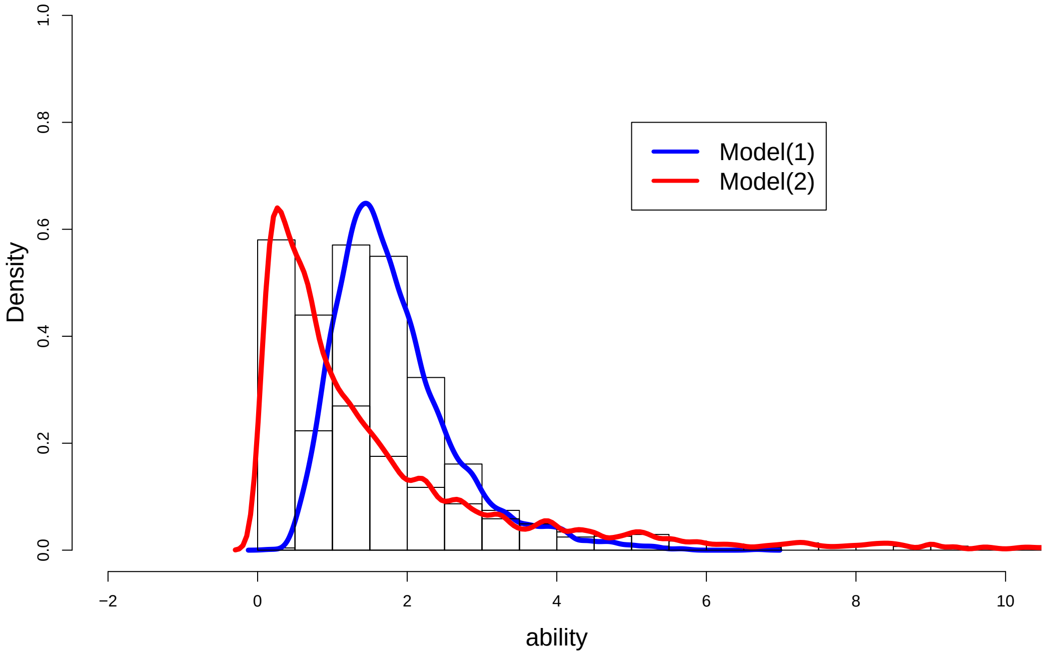

Figure C.1: Distribution of authors' abilities from Model (1) with an assumed exogenous coauthorship network and from Model (2) with endogenous author-project matching.

showing that Model (1) generates upward biases on the estimates of research effort. Finally, we plot the distributions of predicted project output in Figure C.3. From the figure, it is clear that Model (1) (blue line) fits the real data (green line) much worse than Model (2) (red line). Therefore, ignoring author- and project-specific effects makes Model (1) less suitable for the paper output data.

Next, we should answer why the estimates of λ and ϕ are biased downward in Model (1). Our argument is, because Model (1) overestimates author ability and research effort as shown in Figures C.1 and C.2, a smaller (or even negative) collaborative spillover parameter is estimated to fit the level of project outputs observed in the data. Implicatively, a higher collaborative spillover parameter in Model (1) could potentially lead to an over-prediction of project outputs. To support this argument we re-estimate Model (1) with restrictions as in the following three counterfactual scenarios: (i) fixing λ at the value from Model (2) (ii) fixing ϕ at the value from Model (2); (iii) fixing both λ and ϕ at the values from Model (2) in Table 2. Then we use these counterfactual estimation results to predict project outputs, which are shown in Figure C.4. From the figure we observe that none of these restricted models provides a better fit to the data than the original unrestricted Model (1); and thus, none of them would achieve a higher likelihood value compared with the original Model (1) without fixing any parameter value. This shows that higher values of λ and ϕ would not be supported from estimation of Model (1). Thus, compared with Model (2), which incorporates author- and project-specific effects to solve the potential omitted variables problem, the estimation results of λ and ϕ in Model (1) are

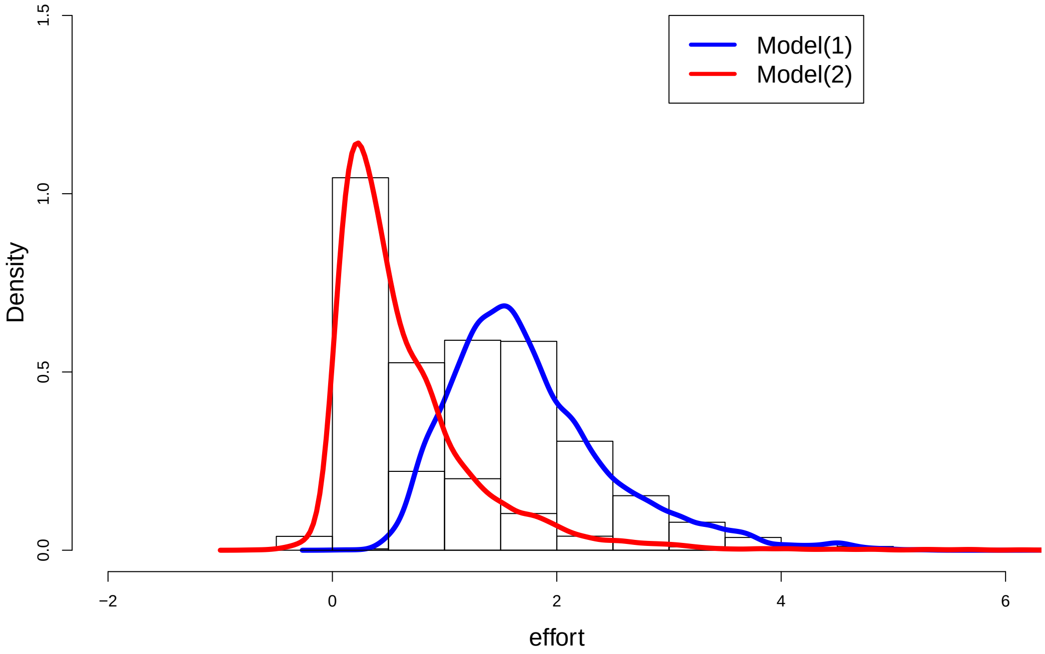

Figure C.2: Distribution of effort levels computed from Model (1) with an assumed exogenous coauthorship network and Model (2) with endogenous author-project matching.

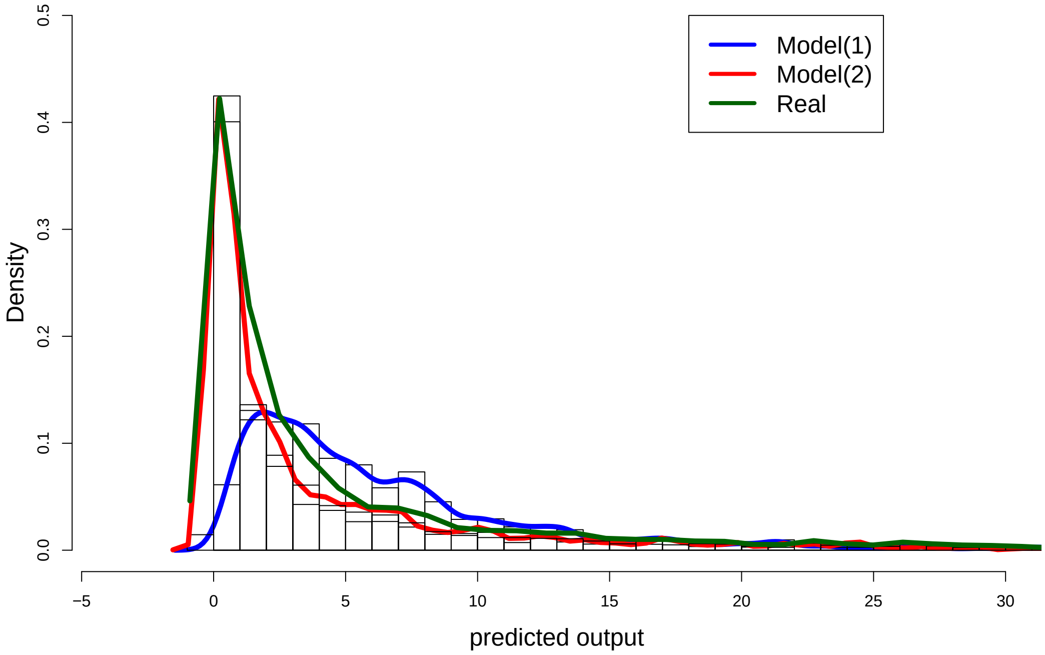

Figure C.3: Distribution of papers' outputs computed from Model (1) with an assumed exogenous coauthorship network; from Model (2) with endogenous author-project matching; and from the data.

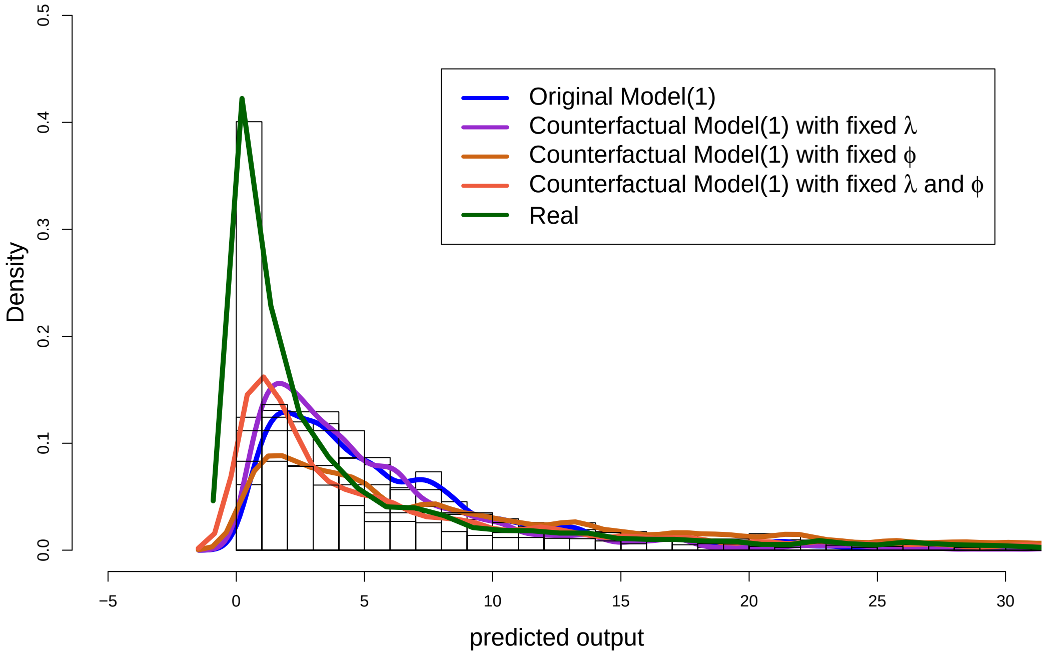

Figure C.4: Distribution of papers' qualities computed from Model (1) with an assumed exogenous coauthor network; from counterfactual Model (1) with fixed λ; from counterfactual Model (1) with fixed ϕ; from counterfactual Model (1) with fixed λ and ϕ; and from the data (green).

downward biased.

D. Performance of the Bayesian MCMC Estimation Approach

To show that the Bayesian MCMC estimation approach in Section 4.4 can effectively recover the true parameters from the model of Equations (15) and (16), we conduct a Monte Carlo simulation to study the bias and standard deviation from the estimation results. The simulation consists of 100 repetitions. In each repetition, the data-generating process (DGP) runs as follows: we first simulate dyadic binary exogenous variables $z_{is} \in \{0, 1\}$ by drawing two uniform random variables, u_i and u_s. If both u_i and u_s are above 0.7 or below 0.3, we set $z_{is} = 1$; otherwise, we set $z_{is} = 0$. We simulate individual exogenous variables $\mathbf{x}$, author latent variables $\boldsymbol{\mu}$, and project latent variables $\boldsymbol{\kappa}$ from standard normal distributions. Then, we generate the artificial project participation $\mathbf{G}$ and project output $\mathbf{Y}$ based on the matching function of Equation (15) and the production function of Equation (16). After obtaining the artificial data, we estimate two models: one is the true model of DGP where both project output and project participation are endogenous and the other is just the production function by treating the participation matrix $\mathbf{G}$ as exogenous. We conduct simulations with two sample sizes to show how data information can improve estimation accuracy in finite samples.

The simulation results are summarized in Table D.1. We report the bias and the standard

Table D.1: Simulation results.

		n=200, p=250				n=300, p=350			
		Exo. Net.		Endo. Net.		Exo. Net.		Endo. Net.	
	DGP	Bias	S.D.	Bias	S.D.	Bias	S.D.	Bias	S.D.
λ	0.0500	-0.0311	0.0236	-0.0023	0.0066	-0.0226	0.0109	-0.0007	0.0029
ϕ	0.0500	-0.1102	0.0186	0.0106	0.0314	-0.0858	0.0101	-0.0005	0.0130
β_1	0.5000	1.3659	0.2813	-0.2559	0.1526	1.3644	0.1573	-0.1354	0.0783
β_2	0.5000	-0.1667	0.1384	0.0023	0.0580	-0.1819	0.0697	0.0021	0.0308
ζ	2.0000			0.0789	0.1629			0.0337	0.0869
η	0.5000			0.2413	0.1614			0.1039	0.1162
σ^2	1.0000	27.7923	11.0436	-0.1908	0.1394	37.6195	10.5491	-0.1610	0.0958
γ_{10}	-5.5000			-0.2344	0.1209			-0.0943	0.0832
γ_{11}	0.5000			-0.0428	0.1486			0.0010	0.0917
γ_2	1.0000			0.0650	0.0751			0.0337	0.0536
γ_3	0.5000			0.1996	0.0899			0.0867	0.0575

deviation based on the point estimate of each coefficient across repetitions. First of all, we observe that when treating the collaboration network as exogenous, there are downward biases on the estimates of λ and ϕ. This mimics the problem that we saw from the empirical study, which reassures our argument that omitting individual latent variables would overestimate the authors' abilities, which results in lower estimates of λ and ϕ. The second thing to be observed from the table is, when using the full model, we mostly recover the true value of each coefficient, despite the small sample biases. However, these finite sample biases fade away when the sample size increases, which indicates that the proposed estimation algorithm has the desired finite sample performance.

E. Goodness-of-Fit Statistics

The matching model outlined in Section 4.3 attempts to uncover a channel in which authors choose projects in which to participate. Based upon participation, authors form coauthorship links with others. A way to tell whether this matching model explains the real data well or not is to conduct a goodness-of-fit examination for the implied coauthor network.

We follow Hunter et al. [2008] to conduct the goodness-of-fit examination. We take the observed coauthor network data from the real sample. Then we simulate one hundred artificial networks from our matching model with parameters reported in Table 2. Model fitness is examined by the similarity between simulated networks and observed networks in the distribution of four network statistics – degree, edge-wise shared partner, minimum geodesic distance, and average nearest neighbor connectivity.

In order to simulate artificial coauthorship networks, we follow the iteration approach of Snijders [2002]. In this approach, the simulated bipartite collaboration network $\mathbf{G}$ at different iterations t, $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots, \mathbf{G}^{(t)}$, form a Markov chain and the transition probability of the Markov chain is given by

$$\mathbb{P}(\mathbf{G}^a, \mathbf{G}^b) = \mathbb{P}(\mathbf{G}^{(t+1)} = \mathbf{G}^b | \mathbf{G}^{(t)} = \mathbf{G}^a),$$

for $\mathbf{G}^a, \mathbf{G}^b \in \mathcal{G}(n,p)$, where $\mathcal{G}(n,p)$ denotes the set of all bipartite collaboration network matrices $\mathbf{G}$ with the same number n of authors and p projects. Following Equation (17) of Snijders [2002], we simulate $\mathbf{G}$ from the transition probability by the Meteropolis-Hastings (M-H) algorithm: at each iteration, we randomly choose an element δ_{is} from $\mathbf{G}^{(t)}$ and change it from $\delta_{is}^{(t)}$ to $1-\delta_{is}^{(t)}$. This change will be accepted by probability

$$\mathbb{P}\left(\delta_{is}^{(t+1)} = 1-\delta_{is}^{(t)} \middle| \mathbf{G}^{(t)}\right) = \min\left\{1, \exp((1-2\delta_{is}^{(t)})\psi_{is})\right\},$$

with function ψ_{is} given in Equation (15) and its estimation result in Table 2. This M-H sampling procedure satisfies the detailed balance condition so that after convergence we can regard the realized $\mathbf{G}$ from the last iteration as the one drawn from its stationary distribution. In practice, we set the number of iterations to $2np$, where n is the number of authors and p is the number of projects. After getting the simulated participation incidence matrix $\mathbf{G}$, we do a projection (cf. Figure 1) to obtain the coauthor network adjacency matrix.

The examination results are shown in Figure F.1. We present the distributions of statistics for the observed network by solid lines and the distributions for the simulated networks by dotted lines (mean with $5^{th}$and 95^{th} percentiles). As network statistics we consider the degree distribution, edge-wise shared partners, average nearest neighbor degree and the clustering degree distribution.[43] From the figure we find that the simulated networks and the observed network display similar distributions over these four statistics. This suggests that our estimated model is able to emulate the unobserved network-generating process.

F. Additional Robustness Checks

In this appendix we report the various results of our robustness checks to analyze the sensitivity of the estimates shown in Table 2. First, in Table F.1 we show the estimation results using only the similarities in the research (NEP) fields for the matching score of Equation (15). The estimated spillover and congestion effects are similar to the ones reported in Table 2, reassuring that identification comes from the exogenous variation in the research overlap between author and projects. Secondly, Table F.4 shows the estimation results with an alternative paper output measure. While in Table 2 we used the sum of citation recursive impact factors to measure a papers' output, in the first two columns of Table F.4 we simply use the sum of the citation impact factor as an alternative. Then, the next four columns of Table F.4 show the estimation results with alternative sample periods, covering the years 2007 to 2009 and the years 2013 to 2015. The corresponding summary statistics for these two sample periods are reported in Tables F.2 and F.3, resepctively. Comparing summary statistics of three different sample periods in

[43] The edge-wise shared partners contain information of a network related to the count of triangles in a network G. Its distribution consists of values $EP_G(0)/E_G, \cdots, EP_G(m-2)/E_G$, where $EP_G(k)$ denotes the number of edges whose endpoints both share edges with exactly k other nodes and E_G is the total number of edges in network G. The average nearest neighbor connectivity is the average degree of the neighbors of a node. The clustering coefficient measures the fraction of connected neighbors of a node with degree k.

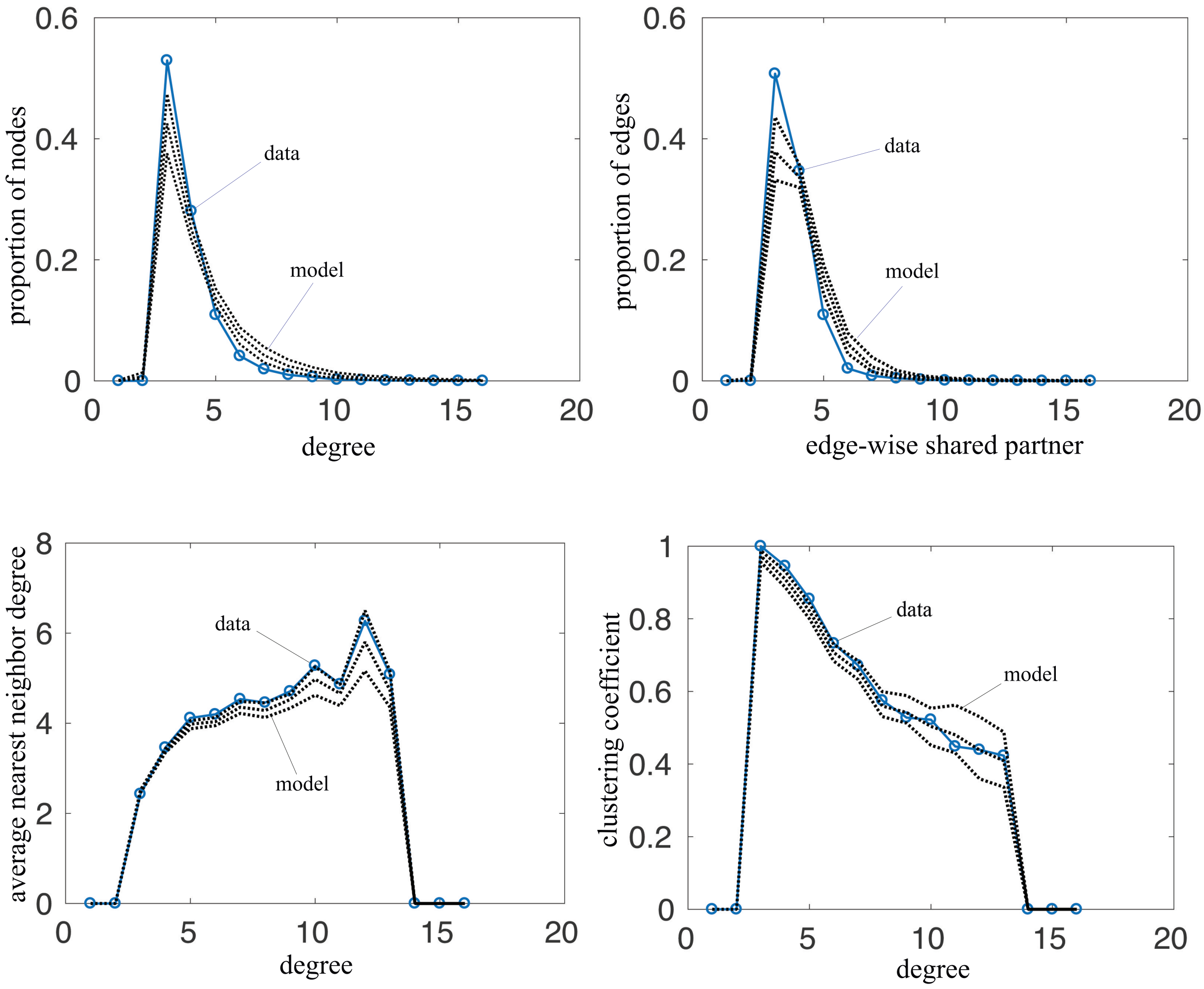

Figure F.1: Goodness-of-fit statistics for the coauthorship network.

Tables 1, F.2 and F.3, it is clear that the paper output measure declines when the sample period approaches closer to the end date, which is due to a shorter time for papers to accumulate their citations. We also see the average number of co-authors in each paper increases gradually over time from 1.576 (in years 2007-2009) to 1.661 (in years 2013-2015). From the author side, most author attributes remain similar over time, despite that authors in the earlier sample have more lifetime citations and longer years of experiences.

Estimation results are presented in Table 2 for the exogenous network (Exo. Net.) and the endogenous network (Endo. Net.) cases. Similar to the results of Table 2 in the main text, the estimates of λ and ϕ in the exogenous network case are downward biased due to omitting the endogenous matching between authors and projects; and the biases can be corrected when we jointly model paper output and the formation of the coauthor network. Except for the same pattern of bias correction, the results further show that the spillover and congestion effects rise over time across sample periods, implying the increasing importance of the coauthor network on economic research.

Table F.1: Estimation results for the 2010-2012 sample with matching based only on same NEP fields

		Homogeneous Spillovers (1)	Heterogeneous Spillovers (2)	Discounting # of Coauthors (3)
Output				
Spillover	(λ)	0.0152***	0.0210**	0.0423**
		(0.0056)	(0.0103)	(0.0209)
Cost	(ϕ)	0.9118***	0.9415***	0.9535***
		(0.0295)	(0.0421)	(0.0310)
Constant	(β_0)	-1.1911***	-1.2343***	-1.2027***
		(0.1273)	(0.1227)	(0.1259)
Log life-time citat.	(β_1)	0.3024***	0.3091***	0.3093***
		(0.0175)	(0.0178)	(0.0160)
Decades after grad.	(β_2)	-0.2608***	-0.2280***	-0.2423***
		(0.0222)	(0.0201)	(0.0209)
Male	(β_3)	0.0551	0.0551	-0.0212
		(0.0504)	(0.0414)	(0.0484)
NBER connection	(β_4)	0.281***	0.2634***	0.2636***
		(0.0295)	(0.0282)	(0.0262)
Ivy League connect.	(β_5)	0.3898**	0.3779***	0.3670***
		(0.0304)	(0.0255)	(0.0262)
Editor	(β_6)	-0.3680***	-0.3308***	-0.3117***
		(0.0479)	(0.0424)	(0.0500)
Author effect	(ζ)	1.6419***	1.4648***	1.3649***
		(0.0442)	(0.0399)	(0.0344)
Project effect	(η)	1.7662***	0.9721**	1.1499***
		(0.5753)	(0.4862)	(0.4739)
Project variance	(σ_v^2)	133.0565***	128.9216***	123.8766***
		(2.5848)	(2.4736)	(2.3769)
Matching				
Constant	(γ_0)	-7.7191***	-7.7034***	-7.7301***
		(0.0232)	(0.0243)	(0.0236)
Same NEP	(γ_1)	0.1652*	0.206**	0.1738*
		(0.1040)	(0.1067)	(0.1002)
Author effect	(γ_2)	1.4738***	1.2759***	1.2188***
		(0.0409)	(0.0367)	(0.0314)
Project effect	(γ_3)	-0.0491	-0.0625	-0.0027
		(0.0744)	(0.0642)	(0.0607)
Sample size (papers)		5,587	5,587	5,587
Sample size (authors)		2,930	2,930	2,930

Notes: The dependent variables are project output following Equation (16) and project-author matching following Equation (14) using only the similarities in the research (NEP) fields for the matching score function. Model (1) assumes homogeneous spillovers between coauthors. Model (2) allows for heterogeneous spillovers using Jaffe's similarity measure for the research fields of collaborating authors. Model (3) considers the case where in the utility of an author in Equation (2) we discount the number of coauthors in each project. We implement MCMC sampling for 30,000 iterations and leave the first 1000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (put into the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

Table F.2: Summary statistics for the 2007-2009 sample.

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive discounted impact factor	0.0001	679.0147	10.8133	26.9650	5832
number of authors (in each paper)	1	5	1.5758	0.7080	5832
Authors					
Log life-time citations	0	10.6683	5.6694	1.6648	2812
Decades after Ph.D. graduation	-0.6	7.000	1.2753	1.0184	2812
Male	0	1	0.8105	0.3919	2812
NBER connection	0	1	0.1024	0.3032	2812
Ivy League connection	0	1	0.1501	0.3571	2812
Editor	0	1	0.0562	0.2303	2812
number of papers (for each author)	1	52	3.2681	3.5182	2812

Notes: We drop papers without any citations when extracting from the RePEc database. Authors who only work on these dropped papers are also dropped.

Table F.3: Summary statistics for the 2013-2015 sample.

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive discounted impact factor	0.0001	185.9120	3.5793	7.4761	3575
number of authors (in each paper)	1	5	1.6607	0.7294	3575
Authors					
Log life-time citations	0	10.6683	5.2670	1.9185	2189
Decades after Ph.D. graduation	-0.6	5.400	0.9494	1.0791	2189
Male	0	1	0.8136	0.3895	2189
NBER connection	0	1	0.1023	0.3032	2189
Ivy League connection	0	1	0.1389	0.3459	2189
Editor	0	1	0.0521	0.2222	2189
number of papers (for each author)	1	57	2.7122	3.0181	2189

Notes: We drop papers without any citations when extracting from the RePEc database. Authors who only work on these dropped papers are also dropped.

Table F.4: Estimation results of the sum of simple citation impact factors for the 2010-2012 sample, and alternative sample periods of the years 2007-2009 and the years 2013-2015.

		Alternative Output		2007-2009 Sample		2013-2015 Sample	
		Exo. Net.	Endo. Net.	Exo. Net.	Endo. Net.	Exo. Net.	Endo. Net.
		(1)	(2)	(1)	(2)	(1)	(2)
Output							
Spillover	(λ)	-0.0657**	0.0655***	-0.1040***	0.0169*	-0.0083	0.1071***
		(0.0310)	(0.0200)	(0.0276)	(0.0105)	(0.0296)	(0.0325)
Congestion	(ϕ)	0.0209***	0.3150***	-0.0127***	0.5607***	-0.0046	1.1528***
		(0.0087)	(0.0386)	(0.0021)	(0.0675)	(0.0066)	(0.0458)
Constant	(β_0)	0.7757***	-1.3299***	-0.8202***	-3.8333***	-1.3413***	-3.8467***
		(0.1473)	(0.1717)	(0.1417)	(0.2406)	(0.1861)	(0.2217)
Log life-time citat.	(β_1)	0.2738***	0.5353***	0.3481***	0.7397***	0.3161***	0.5947***
		(0.0208)	(0.0241)	(0.0220)	(0.0311)	(0.0281)	(0.0268)
Decades after grad.	(β_2)	-0.1336***	-0.3487***	-0.3183***	-0.4804***	-0.4040***	-0.3772***
		(0.0344)	(0.0318)	(0.0354)	(0.0339)	(0.0597)	(0.0340)
Male	(β_3)	-0.0891	0.1379***	-0.3077***	0.1869***	0.0552	0.6634***
		(0.0693)	(0.0440)	(0.0430)	(0.0442)	(0.0998)	(0.0645)
NBER connection	(β_4)	0.1175**	0.3336***	0.0649	0.3186***	0.1676***	0.0470
		(0.0552)	(0.0331)	(0.0424)	(0.0364)	(0.0588)	(0.0406)
Ivy League connect.	(β_5)	0.2563***	0.2360***	0.3523***	0.2297***	0.0766	-0.2148***
		(0.0487)	(0.0335)	(0.0373)	(0.0328)	(0.0565)	(0.0447)
Editor	(β_6)	-0.0623	-0.1680***	-0.4148***	-0.3762***	0.0668	-0.2373***
		(0.0731)	(0.0605)	(0.1154)	(0.0719)	(0.0918)	(0.0814)
Author effect	(ζ)	–	1.4787***	–	1.9647***	–	2.9620***
			(0.0532)		(0.0829)		(0.0997)
Project effect	(η)	–	0.2944	–	5.4234***	–	2.1976***
			(1.0078)		(0.6266)		(0.2333)
Project variance	(σ_v^2)	151,540***	86,942***	567.0469***	437.3041***	46.2065***	24.7691***
		(2877.9)	(1681.3)	(10.6025)	(8.2949)	(1.1065)	(0.6183)
Matching							
Constant	(γ_0)	–	-14.8468***	–	-22.1408***	–	-19.8048***
			(0.1669)		(0.2786)		(0.3216)
Same NEP	(γ_{11})	–	0.7544***	–	3.4423***	–	4.4419***
			(0.2089)		(0.1086)		(0.1611)
Ethnicity	(γ_{12})	–	5.3876***	–	8.9120***	–	7.5462***
			(0.0896)		(0.1538)		(0.1712)
Affiliation	(γ_{13})	–	5.3960***	–	8.7770***	–	8.1511***
			(0.2784)		(0.3355)		(0.3709)
Gender	(γ_{14})	–	2.7515***	–	4.9347***	–	4.0535***
			(0.1070)		(0.1478)		(0.1689)
Advisor-advisee	(γ_{15})	–	7.3246***	–	9.7784***	–	12.0720***
			(0.1893)		(0.2382)		(0.3089)
Past coauthors	(γ_{16})	–	6.5646***	–	7.3005***	–	6.7537***
			(0.1261)		(0.1857)		(0.2349)
Common co-authors	(γ_{17})	–	10.0296***	–	15.1485***	–	13.5523***
			(0.1309)		(0.1707)		(0.2182)
Author effect	(γ_2)	–	2.0381***	–	3.8427***	–	3.6080***
			(0.0539)		(0.0776)		(0.0967)
Project effect	(γ_3)	–	-9.4587***	–	-9.5118***	–	-7.5379***
			(0.1927)		(0.1664)		(0.1814)
Sample size (papers)		5,587		5,832		3,575	
Sample size (authors)		2,930		2,812		2,189	

Notes: The mean, s.d., max, and min the of the sum of the citation impact factors are (164.4077, 422.3750, 13093, 0.0096). Model (1) studies project output of Equation (13) assuming exogenous matching between authors and papers. Model (2) studies project output of Equation (16) assuming endogenous matching by Equation (14). We implement MCMC sampling for 30,000 iterations and leave the first 1000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (put into the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.