Superstar Economists: Coauthorship networks and research output[☆]

Chih-Sheng Hsieh[a], Michael D. König[b], Xiaodong Liu[c], Christian Zimmermann[d]

[a] *Department of Economics, Chinese University of Hong Kong, CUHK Shatin, Hong Kong, China.*
[b] *Department of Economics, University of Zurich, Schönberggasse 1, CH-8001 Zurich, Switzerland.*
[c] *Department of Economics, University of Colorado Boulder, Boulder, Colorado 80309-0256, United States.*
[d] *Department of Economic Research, Federal Reserve Bank of St. Louis, St. Louis MO 63166-0442, United States.*

Abstract

We study the impact of research collaborations in coauthorship networks on total research output. 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 scientific coauthorship network of 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 explicitly takes into account the spillover effect generated in the coauthorship network. Moreover, we analyze various funding instruments for individual researchers as well as their departments, and compare them to the economics funding program by 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.

Keywords: coauthor networks, scientific collaboration, spillovers, key player, research funding, economics of science
JEL: C72, D85, D43, L14, Z13

1. Introduction

We build a micro-founded model for scientific knowledge production networks that incorporates and generalizes previous ones in the literature (cf. e.g. Ballester et al., 2006; Cabrales et al., 2011; Jackson and Wolinsky, 1996). We characterize the equilibrium when multiple agents 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 the total research output, and design the optimal network-based research funding programs.

[☆] We would like to thank Bryan Graham, Lorenzo Ductor, Bauke Visser, Kenan Huremovic, Marco Van Der Leij, Matt Jackson, David Miller, Joachim Voth, Francis Bloch, Pietro Biroli, Ralph Ossa, Hannes Schwandt, Fabrizio Zilibotti and various seminar participants for the helpful comments. Moreover, we would like to thank Adrian Etter and Marc Biedermann for the excellent research assistance. Michael D. König acknowledges financial support from 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.

Email addresses: cshsieh@cuhk.edu.hk (Chih-Sheng Hsieh), michael.koenig@econ.uzh.ch (Michael D. König), xiaodong.liu@colorado.edu (Xiaodong Liu), zimmermann@stlouisfed.org (Christian Zimmermann)

Based on the economic micro-foundation, we develop an econometric model with the unobservable effort levels determined by the Nash equilibrium and the self-selection of agents into different projects determined by a matching process that depends on both the agents' and projects' characteristics (cf. e.g. Chandrasekhar and Jackson, 2012). We estimate this model using data for the network of scientific coauthorships between economists registered in the Research Papers in Economics (RePEc) author service.[1]

We then propose a novel ranking measure for economists and their departments which is derived from the 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).[2] Taking into account endogenous effort choices of the authors, and spillovers generated across the coauthorship network, we find that the highest ranked authors are not necessarily the ones with the largest number of citations, or coincide with other author ranking measures used in the literature. This discrepancy is not surprising, as traditional rankings are typically not derived from microeconomic 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 (see also 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 and the number of lifetime citations of an author. In contrast, the NSF awards are not correlated with the degree and positively but not significantly correlated with the optimal funding policy.

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 scien-

[1] 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 1900 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 database.

[2] Note that the effect of hiring superstar scientists on the profitability of firms has been studied in Hess and Rothaermel (2011); Lacetera et al. (2004); Rothaermel and Hess (2007). In particular, Rothaermel and Hess (2007) define star 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 star scientists takes into account the spillover effects of one scientist on others in a collaboration network.

tific 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, and Ductor (2014); Ductor et al. (2014) study how these collaborations affect research output of individual authors. At an aggregate level, Bosquet and Combes (2013) estimate the effect of department size on its research output. Different to these works, we take a structural approach by introducing a production function for the scientific co-authorship 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),[3] 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 payoff specification, and allow firms to choose endogenously the quantities sold to each market. Different to these authors, the efforts invested by the agents in different projects in our model are strategic complements, and not substitutes as 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, where 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 is 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 (both at the author and the department level) is then provided in Section 5. Section 6 provides the optimal research funding policy and compares it with the economics funding program by the National Science Foundation (NSF). Finally, 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-E.

[3] 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).

2. Theoretical Model

2.1. Production Function

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

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

where Y_s is the research output of project s, e_{is} is the research effort that agent i spent in project s ($e_{is} = 0$ if agent i does not participate in project s), α_i captures the productivity 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).

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 has been used, for example, to analyze production in teams (cf. Adams, 2006).[4]

2.2. Utility Function

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

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

where $g_{is} \in \{0, 1\}$ indicates whether agent i participates in project s, $\delta_s \in (0, 1]$ is a discount factor,[5] and the parameter $\phi > 0$ represents substitutability between the research efforts of the same agent in different projects.[6] This cost is convex if and only if the $p \times p$ matrix $\mathbf{\Phi}$, with the diagonal elements being one and the off-diagonal elements being 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. (2012). Further, a convex separable cost specification can be found in the model studied in Adams (2006).

[4] 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).

[5] 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).

[6] 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.

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

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

where $\otimes$ denotes Kronecker product, $\mathbf{G}$ is an np-dimentional diagonal matrix given by $\mathbf{G} = \mathrm{diag}_{s=1}^{p}\{\mathrm{diag}_{i=1}^{n}\{g_{is}\}\}$, $\mathbf{F}$ 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 being ones. Let $\rho_{\max}(\mathbf{A})$ denotes 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*

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

then the equilibrium effort portfolio is given by

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

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

Observe that the matrix $\mathbf{L}^{\lambda,\phi}$ represents a weighted matrix of the *line graph* $L(\mathcal{G})$ for the bipartite network $\mathcal{G}$,[7] 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 exposition purpose, let $f_{ij} = 1$ for all $i \neq j$ and $\delta_s = 1$ for all s. Following Equation (3),*

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

[7] 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).

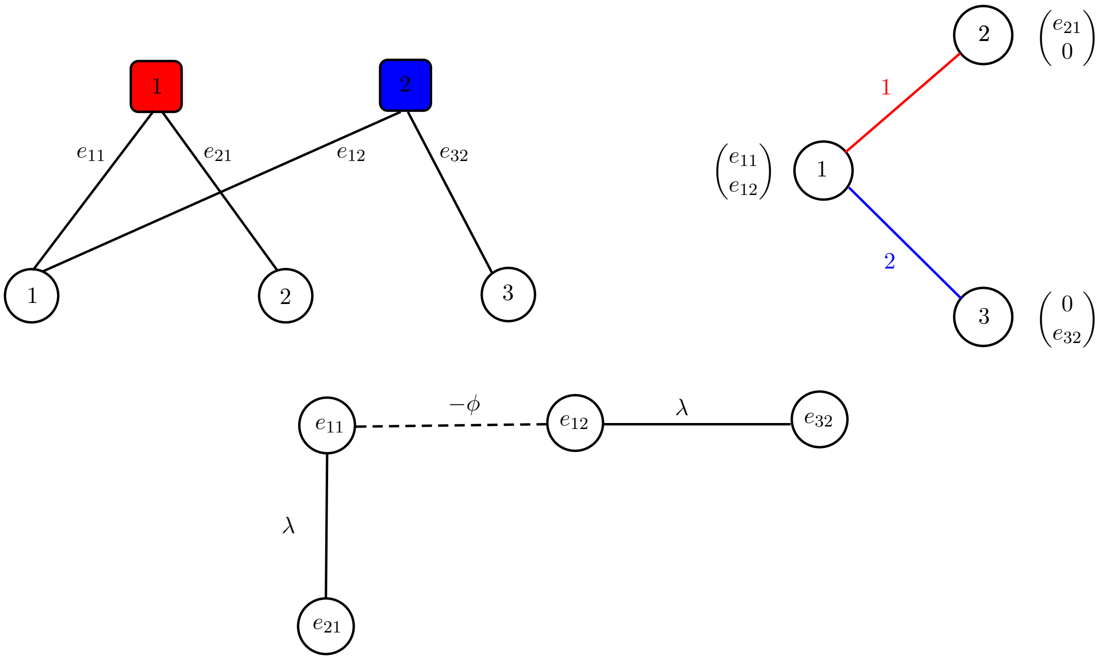


Figure 1: (Top left panel) The bipartite collaboration network $\mathcal{G}$ of authors and projects analyzed in Example 1, where round circles represent authors and squares represent projects. (Top right panel) The projection of the bipartite network $\mathcal{G}$ on the set of coauthors. The effort 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.

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 collaboration partners. 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 $\alpha_2 = 0.5$, $\alpha_3 = 0.8$, $\lambda = 0.1$, $\phi = 0.25$ 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.

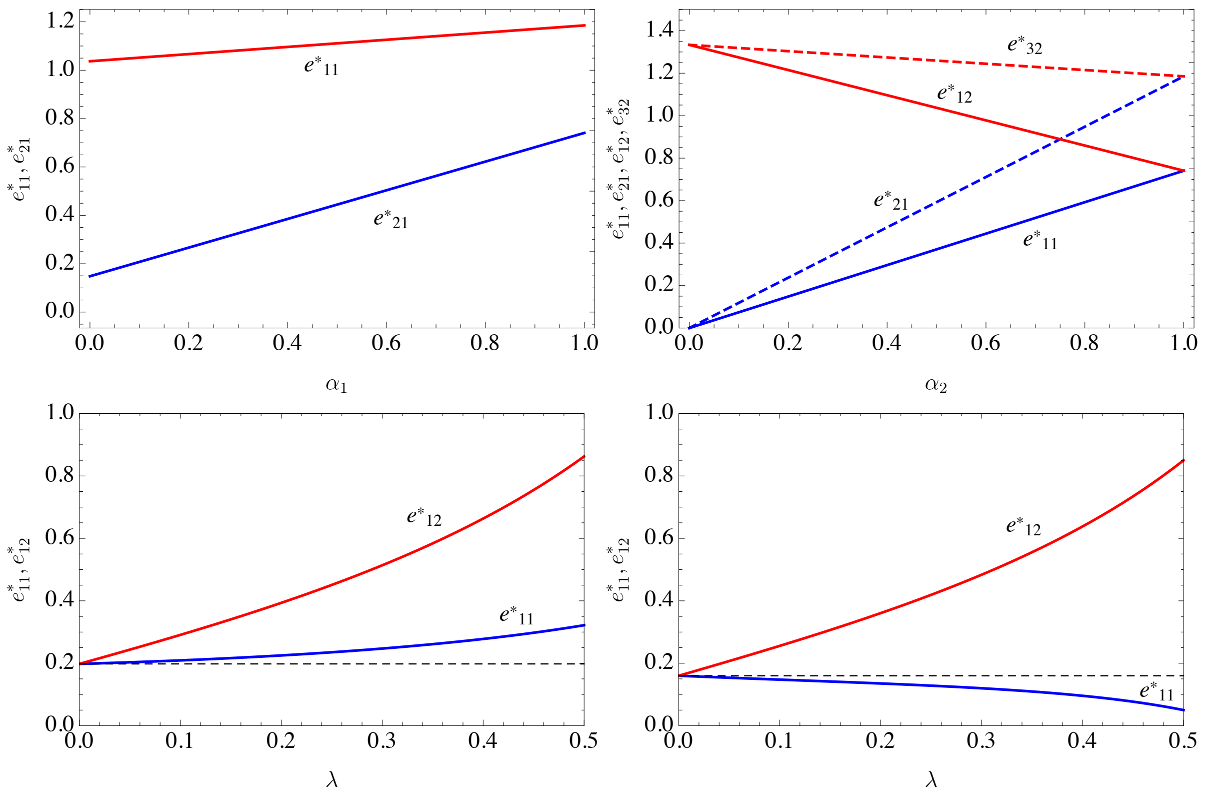


Figure 2: (Top left panel) Equilibrium effort levels for agents 1 and 2 in project 1 for $\phi = 0.75$, $\lambda = 0.25$, $\alpha_2 = \alpha_3 = 1$ (where $e^*_{11} = e^*_{12}$ and $e^*_{21} = e^*_{32}$) and varying values of α_1. (Top right panel) Equilibrium effort levels for agents 1, 2 and 3 in projects 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 $\lambda = 0$.

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 the 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[8]

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

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 the research efforts of collaborating researchers 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.[9] 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, $r \in \mathbb{R}_+$, per unit of the output she generates. Then the utility function (2) can be extended to

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

[8] 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 another. This could guide the academic wage bargaining process when professors get an offer from a competing university.

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

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

Proposition 2. *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 (8). Given the bipartite network $\mathcal{G}$, if*

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

then the equilibrium effort portfolio is given by

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

where $\mathbf{L}_r^{\lambda,\phi} = \lambda(1+r)\mathbf{W} - \phi\mathbf{M}$, $\delta = [\delta_1, \cdots, \delta_p]'$ and $\alpha = [\alpha_1, \cdots, \alpha_n]'$.

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

$$r^* = \underset{r\in\mathbb{R}_+}{\operatorname{argmax}} \sum_{s\in\mathcal{P}} \left(Y_s(\mathcal{G}, r) - r\sum_{i\in\mathcal{N}} g_{is}\delta_s Y_s(\mathcal{G}, r) \right), \tag{11}$$

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

4. Empirical Implications

4.1. Data

The data used for this study makes extensive use of the metadata assembled by the RePEc initiative and its various projects. RePEc assembles the information about publications relevant to economics from close to 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 2.4 million records, including 0.75 million for pre-prints.[10]

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 (51,000 authors) which include what they have published and where they are affiliated.[11] Second, we get information about their advisors, students and alma mater, as recorded in the RePEc Genealogy project.[12] This academic genealogy data

[10] See http://repec.org/ for a general description of RePEc.
[11] https://authors.repec.org/
[12] https://genealogy.repec.org/

has been complemented with some of the data used in Colussi (2017). Third, we gather in which mailing lists the papers have been disseminated through the NEP project.[13] The latter have human editors determining to which field new working papers belong. Fourth, we make use of paper download data that is made available by the LogEc project.[14] Fifth, we use citations to the papers and articles as extracted by the CitEc project.[15] Sixth, we use journal impact factors, author and institution rankings from IDEAS.[16] 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.[17]

Compared to 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.[18] We believe that proportion is higher for the younger generation that is more familiar with social networks and online tools and thus more likely to sign up to online services. Note also that the 51,000 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 then some. Almost 3,000 journals are listed, most of them with extensive coverage. References are extracted for about 30% of their articles to compute citation counts and impact factors. The lacking references principally come from publishers refusing to release them as they are considered copyrighted. 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 significantly cited in outlets outside of economics that may or may not be indexed in RePEc (several top management, statistics and political science journals are indexed).

The amount of data that is available for this project is overwhelming for the methods we need to adopt to estimate the model. For this reason, 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 which had a first pre-print version within a given span of years. We chose 2010–2012 because it is old enough to give all authors the chance to have added the paper to their profile and for the paper to have been eventually published in a journal. 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.

[13] https://nep.repec.org/

[14] http://logec.repec.org/

[15] http://citec.repec.org/

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

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

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

2. We require all authors of the papers to be registered with RePEc.
3. We require that we can find in the RePEc Genealogy where all authors studied and under which advisor(s).
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 8,160 papers written by 3,478 distinct authors for which we have complete data. The numbers are similar for other years.[19] In our empirical model, we use the number of citations of the paper weighted by recursive discounted impact factors (IF) of the citing outlet as the measure of a paper's output. Thus, we further drop 601 papers which do not have any citations up to July 2017 when retrieving from RePEc. Although it is natural to apply our methodology to this full sample with both single authored and co-authored papers, computation is a problem remained to be solved. In the current implementation, we mainly focus on a subset of samples which are multi-authored. We call it "co-authored sample" and there are 3,620 papers and 1,925 authors involved. This co-authored sample would be mostly relevant due to the higher intensity of collaborations and potentially higher spillovers across authors. In addition, we extend this co-authored sample to accommodate the single-authored papers written by authors who are involved in the co-authored sample. This extended sample allows us to provide robustness check on results when authors' efforts in their co-authored projects could be diluted by their efforts on separate single-authored projects.

To understand an author's productivity, we use explanatory variables including author's log life time citations (at the point of sample collection), years after his/her Ph.D. graduation, dummy variables for being a female, having the NBER affiliation, and graduating from the Ivy League. The summary statistics of variables that we use in our empirical model are provided in Table 2. A detailed description of variables can be found in Appendix B. Figure 5 shows the highly skewed distributions of authors per paper, the number of papers per author and the paper quality. Moreover, Figure 3 shows the collaboration network among authors and Figure 4 the network of collaborations of departments/institutions from the RePEc database. The network of departments is more concentrated on a few central institutions than the network of coauthors.

4.2. Empirical Production Function

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

$$Y_s = \sum_{i \in \mathcal{N}} \alpha_i e_{is} + \frac{\lambda}{2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N} \setminus \{i\}} f_{ij} e_{is} e_{js} + \epsilon_s, \tag{12}$$

where ϵ_s is a paper-specific random shock. We assume $\alpha_i = \mathbf{x}_i' \boldsymbol{\beta}$, where $\mathbf{x}_i$ is a $k \times 1$ vector of author-specific exogenous characteristics. The empirical production function can be estimated

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

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

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive Impact Factor	0.0000	115.5851	6.5796	12.2021	3620
number of authors (in each paper)	1	5	1.8892	0.7108	3620
Authors					
Log life-time citations	0	10.5516	5.4948	1.7118	1925
Decades after Ph.D. graduation	-0.6	5.9000	1.1113	0.9909	1925
Female	0	1	0.1345	0.3413	1925
NBER connection	0	1	0.1195	0.3244	1925
Ivy League connection	0	1	0.1553	0.3623	1925
Editor	0	1	0.0494	0.2167	1925
number of papers (for each author)	1	74	3.5527	3.8339	1925

Note: We drop authors who did not coauthor with any others during the sample period. We also drop papers without any citations when extracting from the RePEc data base.

Table 2: Summary statistics for the 2010-2012 raw sample.

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive Impact Factor	0.0000	115.5851	3.5230	11.7103	8160
number of authors (in each paper)	1	5	1.4942	0.6849	8160
Authors					
Log life-time citations	0	10.5516	4.9995	1.9249	3478
Decades after Ph.D. graduation	-0.6	5.9000	1.0880	1.0614	3478
Female	0	1	0.1377	0.3447	3478
NBER connection	0	1	0.0900	0.2862	3478
Ivy League connection	0	1	0.1300	0.3363	3478
Editor	0	1	0.0451	0.2076	3478
number of papers (for each author)	1	96	3.5058	4.4183	3478

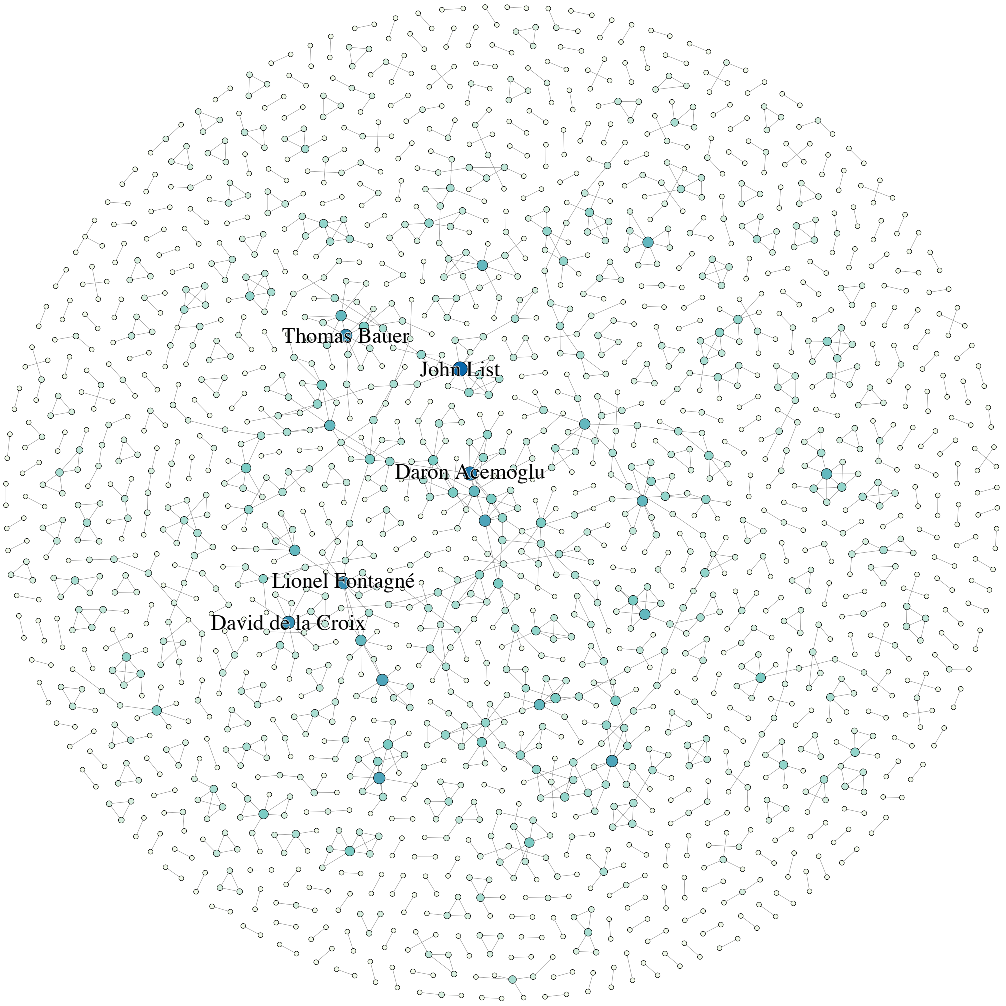


Figure 3: The collaboration network among authors in the RePEc database considering only coauthored projects and dropping projects with zero impact factor. A nodes' size and shade indicates its degree. The names of the five authors with the largest number of coauthors (degree) are indicated in the network.

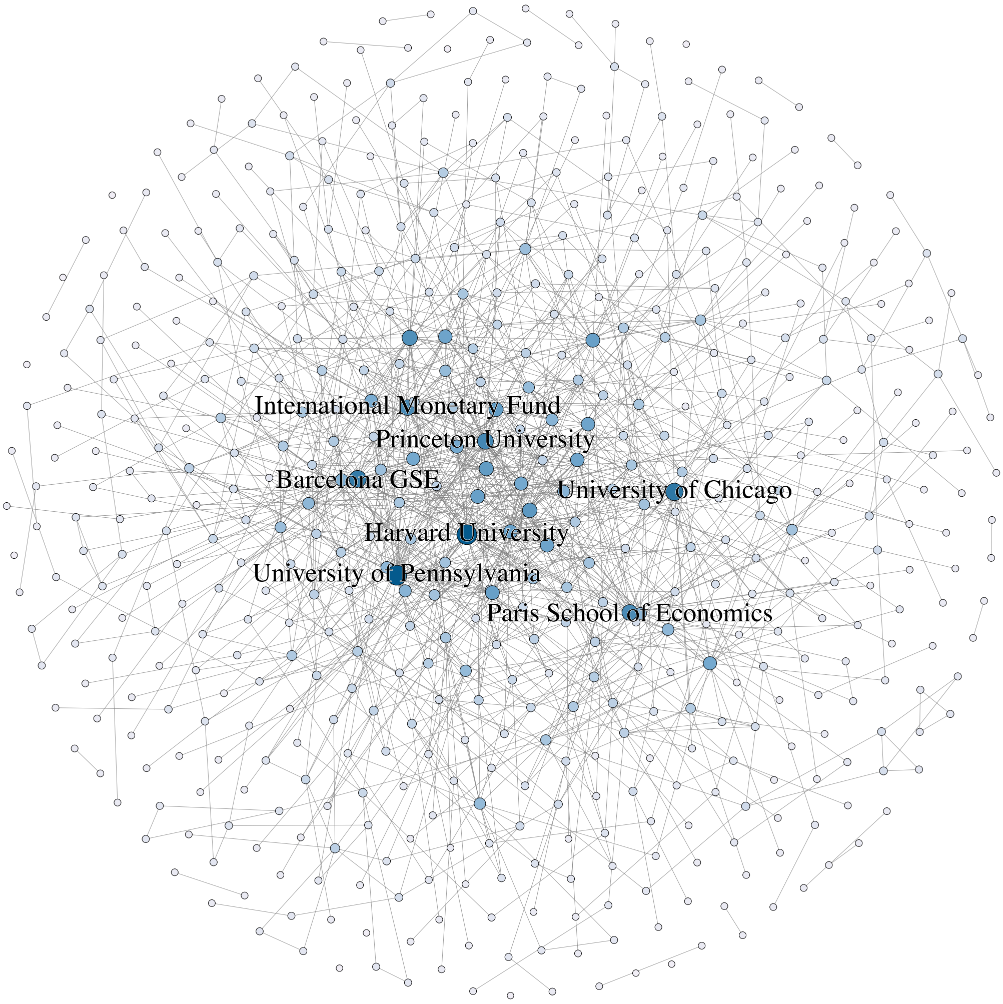

International Monetary Fund
Princeton University
Barcelona GSE
University of Chicago
Harvard University
University of Pennsylvania
Paris School of Economics

Figure 4: The collaboration network among departments in the RePEc database with a total of 867 unique departments. A nodes' size and shade indicates its degree. The names of the five departments with the largest degrees are indicated in the network.

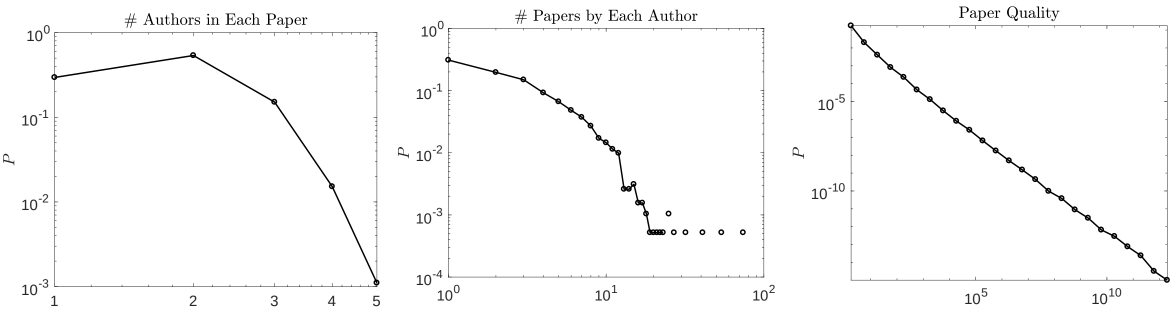


Figure 5: The distribution of authors per paper (left panel), the number of papers per author (middle panel) and the paper quality (right panel).

by the the nonlinear least squares (NLS) method or the maximum likelihood (ML) method (under the normality assumption on ϵ_s), with the unobervable e_{is} replaced by the equilibrium research effort given in Equation (5).

4.3. Matching Process and Identification Strategy

As the equilibrium research effort portfolio given in Equation (5) 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 (12) may suffer from a *self-selection bias* due to the endogeneity of $\mathbf{G}$.

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 the coauthor network $\mathbf{G}$ may suffer from a self-selection bias. To resolve this self-selection bias, we use Heckman's selection-correction approach (cf. e.g. 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 precisely, to address the self-selectivity problem, we model the endogenous matching process of author i to paper s by

$$g_{is} = \mathbb{1}(\psi_{is} + u_{is} > 0), \tag{13}$$

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). In particular, we assume

$$\psi_{is} = \mathbf{z}_{is}'\boldsymbol{\gamma}_1 + \gamma_2\mu_i + \gamma_3\kappa_s, \tag{14}$$

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 measure similarity as the research overlap in the NEP fields of paper s and author i. The identification of the spillover parameter λ in the production function of Equation (12) then comes from the exogenous variation in the research overlap between author i and the potential projects s.

In our empirical analysis, we also consider a specification where we construct additional variables in $\mathbf{z}_{is}$ from the average characteristics of the authors collaborating in project s based on similarity in gender, ethnicity, research field, and whether they have an advisor-advisee relationship (cf. Colussi, 2017). For example, Fafchamps et al. (2010) observe that one of the main determinants of coauthorship is the similarity of research interests between authors. Similarly, Ductor (2014) documents a high assortativity in the matching process in scientific coauthorship networks. By explicitly modeling this assortative matching process through Equation (13) we are able to estimate the causal effect of co-authorship on research output (cf. Ductor, 2014).[20]

The variable μ_i represents author i's unobservable characteristics, and κ_s represents paper

[20] Differently to Ductor (2014), however, our matching equation allows us to control not only for author but also for paper specific effects.

s's unobservable characteristics. The individual fixed effect, μ_i, accounts for all time-invariant unobservable factors, including innate ability, nationality, education, gender, and others. Including μ_i and κ_s allows us to capture the heterogeneity of authors across papers (cf. Graham, 2014, 2015). 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 (13) is to introduce author and paper specific latent variables, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ and $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_p)$, so that unobserved factors contributing to paper output can be controlled for. In other words, the production function of Equation (12) can be extended to

$$Y_s = \sum_{i \in \mathcal{N}} \underbrace{(\mathbf{x}_i' \boldsymbol{\beta} + \zeta \mu_i)}_{\alpha_i} e_{is} + \frac{\lambda}{2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N} \setminus \{i\}} f_{ij} e_{is} e_{js} + \underbrace{\eta \kappa_s + v_s}_{\epsilon_s}, \tag{15}$$

where v_s is assumed to be independent of u_{is} and normally distributed with zero mean and variance σ_v^2. 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

$$\Pr(\mathbf{Y}, \mathbf{G} | \mathbf{X}, \mathbf{Z}) = \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\kappa}} \Pr(\mathbf{Y} | \mathbf{G}, \mathbf{X}, \mathbf{Z}, \mu, \boldsymbol{\kappa}) \Pr(\mathbf{G} | \mathbf{Z}, \boldsymbol{\mu}, \kappa) f(\boldsymbol{\mu}) f(\boldsymbol{\kappa}) d\boldsymbol{\mu} d\boldsymbol{\kappa}, \tag{16}$$

from which we can estimate the parameter vector $\theta = (\lambda, \phi, \boldsymbol{\beta}', \boldsymbol{\gamma}', \eta, \zeta, \sigma_v^2)'$, with $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \gamma_2, \gamma_3)'$.

Observe that the individual and project specific effects, μ_i and κ_s, respectively, both appear in the outcome Equation (15) and the matching Equation (14). 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 (16) this bias can be corrected for.

4.4. Bayesian Estimation

Since the probability function (16) involves a high dimensional integration of latent variables, it is not easy to apply the ML method even when resorting to a simulation approach. As an alternative estimation method, the Bayesian MCMC approach can be more efficient for estimating latent variable models (cf. Zeger and Karim, 1991). We divide the parameter vector θ and other unknown latent variables into blocks and assign the prior distributions as follows:

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

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 $\sigma_\lambda^2 = \sigma_\phi^2 = \sigma_\eta^2 = 10$, $\mathbf{\Xi}_0 = 10\mathbf{I}_{k+1}$, $\boldsymbol{\gamma}_0 = 10\mathbf{I}_{h+2}$, $\tau_0 = 2.2$, and $\nu_0 = 0.1$.

The MCMC sampling combines the Gibbs sampling and the Metropolis-Hastings algorithm, which consists of the following steps:

I. For $i = 1, \cdots, n$, draw the latent variable μ_i using the Metropolis-Hastings algorithm based on $\Pr(\mu_i|\mathbf{Y}, \mathbf{G}, \theta, \mu_{-i}, \kappa)$.

II. draw $\boldsymbol{\gamma}$ using the Metropolis-Hastings algorithm based on $\Pr(\boldsymbol{\gamma}|\mathbf{Y}, \mathbf{G}, \theta \setminus \boldsymbol{\gamma}, \mu, \kappa)$.

III. For $s = 1, \cdots, p$, draw the latent variable κ_s using the Metropolis-Hastings algorithm based on $\Pr(\kappa_s|\mathbf{Y}, \mathbf{G}, \theta, \mu, \kappa_{-s})$.

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

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

VI. draw ξ using the Metropolis-Hastings algorithm based on $\Pr(\xi|\mathbf{Y}, \mathbf{G}, \theta \setminus \xi, \mu, \kappa)$.

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

VIII. draw σ_v^2 using conjugate inverse gamma conditional posterior distributions.

4.5. Estimation Results

We report estimation results for both cases of homogeneous and heterogeneous spillovers in Table 3. In each case, the first column, i.e., Model (1), shows the results where we have assumed that the collaboration network is exogenously given, and the estimation procedure is solely based on the production function outlined in Section 4.2. The second column, i.e., Model (2), allows the collaboration network to be formed endogenously, and is based on the joint estimation of the production function and the matching process in Section 4.3.

First of all, in case of homogeneous spillover, we find in Model (1) that the spillover effect of efforts between co-authors, measured by λ, does not have the expected positive sign. In addition, the congestion effect between projects, measured by ϕ, is insignificant. When comparing to the results in Model (2), where the estimates of λ and ϕ are both significant and having the expected signs, we conclude that the estimates of β and ρ in Model (1) are downward biased due to the problem of endogenous matching between authors and projects. To show why biases are downward, we provide a heuristic explanation in Appendix C by using the estimates from Model (1) and Model (2) to simulate author abilities, efforts, and predicted paper outputs. We show that if the estimates of λ and ϕ in Model (1) become higher than what they currently are, the predicted paper outputs will further deviate from the true ones. 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 also see the same downward biases on the estimates of λ and ϕ when the collaboration network was treated exogenous mistakenly.

Speaking to the effect of author characteristics in paper output, we also find correcting the problem of endogenous matching would change some estimates from Model (1) to Model (2). Based on the results in Model (2), the coefficient of the number of lifetime citations (0.7293) is a positive and significantly predictor (cf. e.g. Ductor, 2014). Experience (measured by decades after Ph.D. graduation) is significantly negative.[21] This finding mirrors Ductor (2014) who shows that career time has a negative impact on productivity and it is argued that this is consistent with the academics' life-cycle effects documented in Levin and Stephan (1991). Female dummy (0.2907) is positively affecting output (cf. Ductor et al., 2017; Krapf et al., 2017). Being affiliated with the NBER (1.1850) positively and significantly impacts research output. Having attended an Ivy League university (0.4761) also positively affects output. The editor dummy has a negative but insignificant effect on output. The author-specific latent variable (4.7197) is found positively and significantly affecting author's productivity. There is no significant effect from paper's latent variable on paper's output.

For the matching between authors and projects, we find that having the same ethnicity, having the same affiliation, being past co-authors, and sharing common co-authors with the lead researcher of a publication make matching more likely (cf. Freeman and Huang, 2015). Similarities in the NEP fields also positively and significantly affect matchings (Ductor, 2014). Being in a Ph.D. advisor–advisee relationship also largely contributes to matchings. Finally, author's latent variable shows a positively significant effect on the author-project matching.

Authors might differ in their competencies and knowledge bases. These differences can effect 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 and then incorporate the proximity measure into the production function of Eq. (1), following Jaffe (1986) for the analysis of technological proximity of patents. In the second case of Table 3 we again find that when omitting the endogenous matching of authors and papers, the estimate of λ and ϕ are downward biased. Also, none of author characteristics show significant coefficients. After coping with the endogenous matching in the full model, the estimate of λ resumes significant and shows a sightly larger value compared to the homogeneous spillover case; meanwhile, the estimate of ϕ also becomes significant but has a slightly smaller value compared to the homogeneous case.

5. Rankings for Individuals and Departments

With our estimates from the previous section (cf. Table 3) we are now able to perform various counterfactual studies. We first investigate the reduction in total output upon the removal of

[21] Following Rauber and Ursprung (2008) and Krapf et al. (2017) 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 (-1.4071) is significantly negative, while the remaining higher orders are insignificant.

Table 3: Estimation results for the 2010-2012 sample.[a]

		Homogeneous Spillovers		Heterogeneous Spillovers		Discounting # Coauthors	
		Model (1)	Model (2)	Model (1)	Model (2)	Model (1)	Model (2)
Output							
Spillover	(λ)	-0.1084***	0.0372*	-0.0954**	0.1679***	-0.1883***	0.2028***
		(0.0340)	(0.0164)	(0.0479)	(0.0229)	(0.0594)	(0.0465)
Cost	(ϕ)	0.0068	0.2696***	0.0057	0.2203***	0.0070	0.2274***
		(0.0054)	(0.0414)	(0.0048)	(0.0138)	(0.0051)	(0.0223)
Constant	(β_0)	-0.6712***	-2.0245***	-0.7263***	-2.0295***	-0.6508***	-2.1625***
		(0.1322)	(0.1681)	(0.1402)	(0.0528)	(0.1304)	(0.0896)
Log life-time citat.	(β_1)	0.2475***	0.4147***	0.2477***	0.3569***	0.2450***	0.3771***
		(0.0227)	(0.0213)	(0.0238)	(0.0125)	(0.0222)	(0.0136)
Decades after grad.[b]	(β_2)	-0.2497***	-0.4207***	-0.2417***	-0.3022***	-0.2463***	-0.3001***
		(0.0400)	(0.0334)	(0.0414)	(0.0222)	(0.0401)	(0.0198)
Female	(β_3)	0.1871***	0.0674**	0.1892***	0.1652***	0.1857***	0.3310***
		(0.0656)	(0.0359)	(0.0681)	(0.0224)	(0.0647)	(0.0352)
NBER connection	(β_4)	0.2779***	0.5564***	0.2784***	0.5308***	0.2773***	0.4195***
		(0.0506)	(0.0407)	(0.0518)	(0.0252)	(0.0520)	(0.0409)
Ivy League connect.	(β_5)	0.1788***	0.1325***	0.1885***	0.3825***	0.1824***	0.1964***
		(0.0455)	(0.0306)	(0.0472)	(0.0356)	(0.0465)	(0.0289)
Editor	(β_6)	-0.1107	-0.3720***	-0.1010	-0.3644***	-0.1088	-0.5073***
		(0.0919)	(0.1391)	(0.0950)	(0.0666)	(0.0915)	(0.0699)
Author effect	(ζ)	–	2.3231***	–	2.4036***	–	2.6691***
			(0.0955)		(0.0307)		(0.2772)
Project effect	(η)	–	0.1305	–	-0.7522	–	-0.4006
			(0.7708)		(0.8211)		(0.7590)
Project variance	(σ_v^2)	118.5337***	76.0442***	119.0462***	76.4599***	118.6615***	73.3228***
		(2.8501)	(1.9560)	(2.8229)	(2.0283)	(2.8154)	(2.0266)
Matching							
Constant	(γ_0)	–	-8.1970***	–	-8.1302***	–	-8.1802***
			(0.0379)		(0.0338)		(0.0349)
Same NEP	(γ_1)	–	5.6924***	–	5.5959***	–	5.6520***
			(0.0658)		(0.0608)		(0.0634)
Author effect	(γ_2)	–	1.8042***	–	1.4971***	–	1.6662***
			(0.1854)		(0.0721)		(0.1882)
Project effect	(γ_3)	–	-0.0581	–	0.0281***	–	0.0622***
			(0.1596)		(0.1698)		(0.1405)
Sample size		3,620		3,620		3,620	

[a] Model (1): Assuming exogenous matching between authors and papers. Model (2): Assuming endogenous matching by Equation (13). The asterisks ***(**,*) indicates that its 99% (95%, 90%) highest posterior density range does not cover zero.

[b] Following Rauber and Ursprung (2008) and Krapf et al. (2017) 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.

Table 4: Estimation results for the 2010-2012 sample with assortative matching.

		Homogeneous	Heterogeneous	Discounting
		Spillovers	Spillovers	# Coauthors
Output				
Spillover	(λ)	0.0259*	0.1041**	0.2072**
		(0.0160)	(0.0354)	(0.0720)
Cost	(ϕ)	0.0886***	0.0718***	0.1053***
		(0.0170)	(0.0150)	(0.0192)
Constant	(β_0)	-1.8936***	-1.7634***	-1.9747***
		(0.1359)	(0.1278)	(0.1118)
Log life-time citations	(β_1)	0.3874***	0.3201***	0.3678***
		(0.0200)	(0.0229)	(0.0241)
Decades after graduation[b]	(β_2)	-0.3832***	-0.3057***	-0.3464***
		(0.0455)	(0.0358)	(0.0781)
Female	(β_3)	0.0958	-0.0750	-0.1199***
		(0.0837)	(0.0973)	(0.0735)
NBER connection	(β_4)	0.3331***	0.5110***	0.4674***
		(0.0541)	(0.0496)	(0.0639)
Ivy League connection	(β_5)	0.3305***	0.3037***	0.2693***
		(0.0445)	(0.0341)	(0.0418)
Editor	(β_6)	-0.2834***	-0.3146***	-0.2579***
		(0.0582)	(0.0930)	(0.0683)
Author effect	(ζ)	2.1649***	2.3144***	2.4938***
		(0.0804)	(0.0692)	(0.1866)
Project effect	(η)	1.0018	-0.1673	0.3557
		(0.8442)	(0.6996)	(0.7823)
Project variance	(σ_v^2)	86.0263***	86.6707***	83.1741***
		(2.2478)	(3.4129)	(2.6043)
Matching				
Constant	(γ_0)	-11.8944***	-12.0136***	-11.9365***
		(0.1898)	(0.1935)	(0.1773)
Same NEP	(γ_{11})	4.9061***	4.9149***	4.8952***
		(0.1106)	(0.1102)	(0.1052)
Ethnicity	(γ_{12})	3.6106***	3.6217***	3.5876***
		(0.0978)	(0.0819)	(0.0846)
Affiliation	(γ_{13})	4.2537***	4.1916***	4.2862***
		(0.2608)	(0.3004)	(0.2650)
Gender	(γ_{14})	1.4575***	1.5632***	1.5032***
		(0.1136)	(0.1175)	(0.1060)
Advisor-advisee	(γ_{15})	4.3542***	4.5338***	4.3964***
		(0.1886)	(0.1876)	(0.1825)
Past coauthors	(γ_{16})	4.7607***	4.7197***	4.7388***
		(0.1289)	(0.1465)	(0.1301)
Share common co-authors	(γ_{17})	6.9945***	6.9835***	6.9914***
		(0.1422)	(0.1259)	(0.1456)
Author effect	(γ_2)	3.1442***	3.0515***	3.0826***
		(0.3001)	(0.2547)	(0.4215)
Project effect	(γ_3)	-0.8640***	-1.0946**	-1.1763**
		(0.3242)	(0.4042)	(0.4364)
Sample size		3,620	3,620	3,620

[a] Model (1): Assuming exogenous matching between authors and papers. Model (2): Assuming endogenous matching by Equation (13). The asterisks ***(**,*) indicates that its 99% (95%, 90%) highest posterior density range does not cover zero.

[b] Following Rauber and Ursprung (2008) and Krapf et al. (2017) we have also estimated a polynomial of order five in decades after Ph.D. graduation. The results show that the coefficient of the first order is significantly negative, while the remaining higher orders are insignificant.

individual authors or entire departments from the network (cf. Section 3.1). Note that when an individual author or a department is removed, the collaboration network will be rewired to reflect the impact. The algorithm of network rewiring follows the network simulation method used in the goodness-of-fit examination in Appendix E.

The ranking of individual authors and departments can be found in Tables 5 and 6, respectively. The key author turns out to be John Van Reenen from the London School of Economics. Our results suggest that without this author total output would be about 1.9% lower (cf. column 6 in Table 5). The second and third highest ranked authors are Alberto Alesina from Harvard University and Gianmarco Ottaviano from the London School of Economics. Their impact on research output is similarly high. The London School of Economics and Harvard University are also among the top five ranked institutions in Table 6. We find that highly ranked authors tend to be more specialized, with an above average concentration in the NEP fields in which their papers are being categorized (from the inverse participation ratio (IPR) in column 10 in Table 5 with an average IPR of 8),[22] and a larger breadth of citing papers across NEP fields (column 9 in Table 5). Highly specialized authors can become authorities in their field (Hackett, 2005). These specialized authors provide crucial inputs to high impact 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. From the ranking of authors we also observe that highly ranked authors tend to have a larger number of papers/projects (column 3), a larger number of citations (column 4) and a higher RePEc rank (column 5). Moreover, betweenness centrality (column 8 in Table 5) and the degree (number of coauthors) are negatively correlated with the ranking, that is, the higher the degree and/or the betweenness centralities, the higher the rank.[23] However, these correlations are not significant at 5% level. Importantly, the above indicators do not yield the same ranking that we obtain based on our model and the data. However, this discrepancy is not surprising, as other rankings are typically not derived from microeconomic foundations, and do not take into account spillover effects generated in scientific knowledge production networks.

[22] See Footnote e in Table 5 for a definition of the inverse participation ratio.

[23] Closeness centrality in column 7 in Table 5 turns out not to be significantly correlated with the ranking.

Table 5: Ranking of the top-twenty five researchers from the 2010-2012 sample.

Rank	Name	Proj.	Citat.	RePEc Rank[a]	Output Loss[b]	Close.[c]	Betw.[c]	NEP Cites[d]	NEP IPR[e]	Organization
1	Van Reenen, John	20	6273	87	-1.91%	4.17	52.57	94.82	21.3983	Centre for Economic Performance, London School of Economics
2	Alesina, Alberto	12	13625	39	-1.78%	4.17	29.47	94.86	18.7128	Department of Economics, Harvard University
3	Ottaviano, Gianmarco	17	4302	220	-1.72%	4.14	39.49	91.69	14.4578	Economics Department, London School of Economics
4	Saez, Emmanuel	14	3930	314	-1.69%	4.61	4.64	91.70	11.1100	Department of Economics, University of California-Berkeley
5	Reinhart, Carmen	12	18358	20	-1.60%	4.67	5.64	93.75	9.07441	Kennedy School of Government, Harvard University
6	Angrist, Joshua	6	8230	53	-1.60%	4.55	9.55	95.87	10.8803	Economics Department, Massachusetts Institute of Technology
7	List, John	27	7741	27	-1.59%	4.12	112.67	95.85	8.72953	Department of Economics, University of Chicago
8	Nunn, Nathan	12	1495	656	-1.55%	4.88	0.54	90.52	11.9138	Department of Economics, Harvard University
9	Bergemann, Dirk	32	1018	951	-1.54%	5.12	2.36	72.28	6.88862	Economics Department, Yale University
10	Pischke, Jorn-Steffen	9	2968	459	-1.54%	4.69	3.12	95.67	13.7850	Centre for Economic Performance, London School of Economics
11	Rogoff, Kenneth	8	21001	8	-1.52%	4.42	10.04	94.81	12.9806	Department of Economics, Harvard University
12	Melitz, Marc	9	6763	145	-1.47%	4.76	1.69	92.73	6.85287	Department of Economics, Harvard University
13	Galor, Oded	12	7663	84	-1.46%	4.86	4.02	91.75	11.2015	Economics Department, Brown University
14	Wacziarg, Romain	8	2660	658	-1.44%	4.75	1.89	92.60	10.8025	School of Management, University of California-Los Angeles
15	Bloom, Nicholas	12	4202	188	-1.36%	4.45	8.55	94.81	15.2667	Department of Economics, Stanford University
16	Morris, Stephen	25	3414	284	-1.35%	4.47	11.07	87.55	6.56941	Department of Economics, Princeton University
17	Wolfers, Justin	15	2786	607	-1.34%	4.71	3.64	93.69	18.9070	Economics Department, University of Michigan
18	Frankel, Jeffrey	41	10765	44	-1.34%	4.41	15.21	93.71	13.0174	Kennedy School of Government, Harvard University
19	Rasul, Imran	11	1447	906	-1.33%	4.57	5.60	85.55	17.5409	Department of Economics, University College London
20	Borjas, George	8	6467	114	-1.32%	4.66	6.70	92.65	8.07620	Kennedy School of Government, Harvard University
21	Eichenbaum, Martin	6	10252	68	-1.29%	4.87	1.61	92.65	8.75977	Department of Economics, Northwestern University
22	Black, Sandra	5	2813	563	-1.29%	4.77	2.45	93.68	9.05840	Department of Economics, University of Texas-Austin
23	Lochner, Lance	12	2085	900	-1.27%	4.88	2.66	86.51	9.36363	Department of Economics, University of Western Ontario
24	Basu, Susanto	3	2488	649	-1.22%	4.66	2.91	89.55	11.1390	Department of Economics, Boston College
25	Demirguc-Kunt, Asli	13	9675	98	-1.18%	4.49	8.10	94.73	15.4880	Economics Research, World Bank Group

[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 researcher i is computed as $\sum_{s=1}^{p} Y_s(G) - \sum_{s=1}^{p} Y_s(G^{-i})$ with the parameter estimates from Table 3. See also Equation (3.1) in Section 6.

[c] 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.

[d] NEP cites measures the breadth of citations across NEP fields. Citation breadth is measured by the number i 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 $x = \mod (i/10 + 2)$ (score listed after decimal point).

[e] To gauge the degree of specialization of an author, we compute the inverse participation ratio (IPR) of the NEP fields in which the papers of an author are announced. Let $\mathbf{x}$ be a vector of indicator variables of these NEP fields. Then we can write $\|\mathbf{x}\|_2^2 / \|\mathbf{x}\|_1^2 = \sum_{i=1}^{n} x_i^2 / \left(\sum_{i=1}^{n} |x_i|\right)^2 = \pi(\mathbf{x})^{-1}$, which is the inverse of the participation ratio $\pi(\mathbf{x})$. The participation ratio $\pi(\mathbf{x})$ measures the number of elements of $\mathbf{x}$ which are dominant. We have that $1 \leq \pi(\mathbf{x}) \leq n$, where a value of $\pi(\mathbf{x}) = n$ corresponds to a fully homogenous case, while $\pi(\mathbf{x}) = 1$ corresponds to a fully concentrated case (note that, if all x_i are identical then $\pi(\mathbf{x}) = n$, while if one x_i is much larger than all others we have $\pi(\mathbf{x}) = 1$).

Table 6: Ranking of the top-ten departments from the 2010-2012 sample.

Rank	Organization	Size	RePEc Rank[a]	Output Loss[b]
1	Department of Economics, Harvard University	23	1	-7.46%
2	Kennedy School of Government, Harvard University	14	16	-4.72%
3	Department of Economics, Princeton University	12	8	-4.28%
4	Economics Department, Massachusetts Institute of Technology	12	5	-3.29%
5	Centre for Economic Performance, London School of Economics	8	71	-3.19%
6	Economics Department, University of Michigan	16	31	-3.17%
7	Booth School of Business, University of Chicago	13	6	-2.79%
8	Department of Economics, University of California-Berkeley	10	10	-2.78%
9	Department of Economics, University of Pennsylvania	11	36	-2.76%
10	Economics Department, Yale University	10	19	-2.70%

[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 3. See also Equation (7) in Section 6.

6. Research Funding for Individuals and Departments

In this section we compare our optimal network-based research funding scheme r^* of Equation (11) in Section 3.2 using the parameter estimates from Section 4.5 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).[24,25]

The economist receiving the largest amount of funds from the NSF is Frank Stafford from the University of Michigan with total funds amounting to 33,471,414 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 436,201 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 95,058,724 U.S. dollars,[26] followed by the University of Michigan with a total of 57,749,679 U.S. dollars. The average funding across organizations from the NSF is $2,831,612$ 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 the optimal funding policy solving the planner’s problem stated in Equation (11) with the estimated parameters from Table 3. 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

[24] See https://www.nsf.gov/awardsearch/.
[25] The data coverage before 1976 is incomplete, and we thus discarded years prior to 1976.
[26] The NBER is ranked 12th according to our network-based optimal funding scheme. See also Table 8.

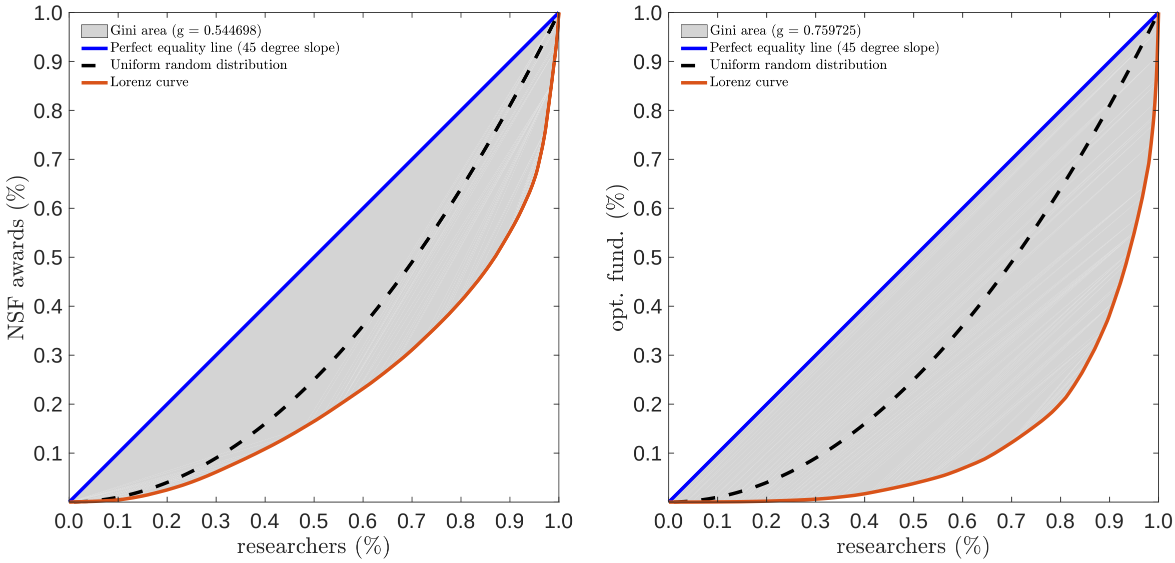


Figure 6: Lorenz curves for the total NSF awards (left panel) and the optimal network-based funding across authors (right panel).

NSF awards and a coefficient of $g = 0.75$ for our network-based optimal funding policy.

Table 7 shows the optimal network-based research funding 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 authors tend to have a larger number of projects and degree/number of coauthors (columns 2 and 3 in Table 7; see also Figure 7), illustrating the importance of the coauthorship network for the optimal funding policy. Moreover, the optimal funding is negatively correlated with closeness centrality, the RePEc rank and positively correlated with the number of citations as well as betweenness centrality (columns 4–7 in Table 7).[27,28] 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.

[27] See also Footnote c in Table 5 for a definition and explanation of the closeness and betweenness centrality measures.

[28] The optimal funding turns out to be uncorrelated with the breadth of citations across NEP fields as measured by NEP cites, and the concentration in the NEP fields measured by the inverse participation ratio (IPR).

Table 7: 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]	NEP IPR[e]	Organization	NSF [%]	Funding [%][f]	Rank
Greg Kaplan	21	5	325	3261	5.0279	1.0236	11.0300	9.8561	Princeton University	0.0875	3.1251	1
Dirk Bergemann	36	5	1018	951	5.1203	2.3644	65.1200	6.8886	Yale University	0.0906	3.0984	2
Nicholas Bloom	13	4	4202	188	4.4500	8.5500	39.1200	15.2668	Stanford University	0.2982	2.7051	3
Olivier Coibion	11	3	765	1699	5.1402	0.7708	76.3600	5.1180	University of Texas-Austin	0.1017	2.5688	4
Fabrizio Perri	11	7	1909	738	4.9014	2.3939	65.1600	7.5378	Federal Reserve Bank of Minneapolis	0.0414	2.4166	5
Stephen Morris	31	4	3414	284	4.4700	11.0700	42.1100	6.5694	Princeton University	0.2152	2.4116	6
Emmanuel Saez	14	7	3930	314	4.6100	4.6400	68.2100	11.1100	University of California-Berkeley	0.2786	2.3734	7
John List	29	12	7741	27	4.1200	112.6700	83.3500	8.7295	University of Chicago	0.0133	2.3509	8
Oded Galor	17	3	7663	84	4.8640	4.0200	37.0900	11.2016	Brown University	0.0822	2.3017	9
Sergio Rebelo	9	4	8043	127	4.9348	1.7689	13.0300	9.1398	Centre for Economic Policy Research	0.0890	2.2698	10
Craig Burnside	10	3	2700	578	5.1033	0.7259	2.0000	9.5135	Duke University	0.0426	2.2153	11
Yuriy Gorodnichenko	8	3	1940	495	4.5500	14.6800	71.2600	14.4722	University of California-Berkeley	0.0839	2.1810	12
Martin Eichenbaum	7	4	10252	68	4.8668	1.6054	89.6000	8.7598	Northwestern University	0.0500	1.9320	13
Vincenzo Quadrini	8	5	1460	1359	5.0292	0.5879	38.1100	9.5144	University of Southern California	0.1836	1.8019	14
Javier Bianchi	9	3	325	3654	5.4083	0.3712	72.2600	6.7236	Federal Reserve Bank of Minneapolis	0.0418	1.7946	15
Joshua Angrist	6	3	8230	53	4.5500	9.5500	92.6100	10.8804	Massachusetts Institute of Technology	0.2597	1.7665	16
Andrei Levchenko	12	6	1081	1120	4.8565	2.4816	95.8100	8.4074	University of Michigan	0.0531	1.7374	17
Sandra Black	5	2	2813	563	4.7700	2.4471	31.0600	9.0584	University of Texas-Austin	0.0588	1.5930	18
Mark Huggett	8	4	1146	1245	5.5324	0.8559	30.0200	7.0588	Georgetown University	0.0128	1.5694	19
John Campbell	5	4	14782	11	4.5900	8.5600	27.0500	11.7769	Harvard University	0.0532	1.5349	20
Chad Syverson	6	4	1656	574	4.7300	4.3500	33.0700	13.8710	University of Chicago	0.0998	1.4443	21
Parag Pathak	3	4	1271	1130	4.8221	2.4116	36.1000	6.7827	National Bureau of Economic Research	0.2258	1.3842	22
Mikhail Golosov	12	9	1077	1025	4.7893	2.3778	3.0000	12.1462	Princeton University	0.0798	1.3701	23
Xavier Gabaix	8	2	3566	185	4.7818	2.5671	72.2500	19.1291	Harvard University	0.1378	1.3505	24
Aleh Tsyvinski	10	7	809	1388	4.8759	1.3933	78.3600	16.1578	Yale University	0.1550	1.2963	25

[a] We only consider 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.

[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 5.

[d] NEP cites measures the breadth of citations across NEP fields. See also Footnote d in Table 5.

[e] The inverse participation ratio (IPR) of the NEP fields meaures the degree of specialization of an author. See also Footnote e in Table 5.

[f] The total cost of funds, $\sum_{s=1}^{p} \delta_{is} r^* Y_s(\mathcal{G}, \mathbf{e}(r^*))$, of researcher i with the optimal research funding scheme r^* of Equation (11) in Section 3.2 with the parameter estimates from Table 3.

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 indicates in bold. The data have been log transformed to account for the heterogeneity across observations.

Further, we find that the rankings of our optimal funding policy and the one by the NSF differ.[29] The author with the highest funds according to our network-based policy is Greg Kaplan from Princeton University (with 3.13% of the total funds) followed by Dirk Bergemann from Yale University (with 3.10% of the total funds) and Nicholas Bloom from Stanford University. However, the third ranked researcher received more than three times as much funding from the NSF as the first ranked researcher. 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.

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).[30] In contrast, the NSF awards are positively but not significantly correlated with the degree or the optimal funding policy.[31] 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 7, but at the departmental level, can be found in Table 8. We find that the Department of Economics at Yale University receives the largest amount of funding, followed by Princeton University. Similar to the ranking of individual authors in Table 7, we observe that the actual funding provided by the NSF does not coincide with the optimal funding policy that we obtain, 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 involved in multiple, possibly overlapping projects. 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 scientific coauthorships 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 the first ranking measure that is based on microeconomic foundations. Moreover, we analyze various funding instruments for individual

[29] The comparison is based on the 236 authors that could be identified in both, the RePEc and the NSF awards databases.

[30] We found no significant correlation of the key player ranking of Table 5 with the optimal funding policy.

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

Table 8: Ranking of optimal research funding for the top-ten departments for the 2010-2012 sample.[a]

Institution	Size	NSF [%]	Funding [%][b]	Rank
Yale University	22	2.8771	8.3996	1
Princeton University	14	2.8250	8.1934	2
Harvard University	46	3.0338	6.7453	3
University of California-Berkeley	24	2.1543	5.9105	4
Federal Reserve Bank of Minneapolis	7	0.2578	5.7235	5
University of Chicago	30	2.6975	4.8246	6
Massachusetts Institute of Technology	18	1.7755	4.6533	7
University of Texas-Austin	12	0.3493	4.5853	8
Stanford University	21	4.0589	4.1962	9
University of Pennsylvania	22	3.0273	4.1644	10

[a] We only consider 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.
[b] The total cost of funds, $\sum_{i \in \mathcal{D}} \sum_{s=1}^{p} \delta_{is} r^* Y_s(\mathcal{G}, \mathbf{e}(r^*))$, for each department $\mathcal{D}$ and researchers $i \in \mathcal{D}$ with the optimal research funding scheme r^* of Equation (11) in Section 3.2 with the parameter estimates from Table 3.

researchers as well as their departments. We show that, because current research 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.

Our analysis can be extended along several directions. First, we can allow the returns of an author from participating in a project to be split equally among the participants of the project similar to the models studied in Jackson and Wolinsky (1996); Kandel and Lazear (1992). Second, instead of a convex cost, we can introduce a time constraint as in Baumann (2014) and Salonen (2016). Third, we can compare our optimal research funding policy with the ones implemented in practice not only in the U.S. but also in other countries. In work in progress we are extending our analysis to the Framework Programs of the E.U. and the research funding program of the Swiss National Science Foundation.

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.

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*, 4(4).

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

Bosquet, C. and Combes, P.-P. (2013). Do large departments make academics more productive? agglomeration and peer effects in research. *CEPR Discussion Paper No. 9401.*

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*, pages 367–383.

Cohen-Cole, E., Liu, X., and Zenou, Y. (2012). Multivariate choice and identification of social interactions. *CEPR Discussion Papers No. 9159*.

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

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*.

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

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):S289–S318.

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. (2014). An empirical model of network formation: detecting homophily when agents are heterogenous. *National Bureau of Economic Research Working Paper No. w20341*.

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

Hackett, E. J. (2005). Essential tensions: Identity, control, and risk in research. *Social studies of science*, 35(5):787–826.

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).

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*, pages 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. *CEPR Discussion Paper No. 9872*.

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.

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. *Advances in strategic management*, 21:133–160.

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

Liu, X. (2014). Identification and efficient estimation of simultaneous equations network models. *Journal of Business & Economic Statistics.*

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.

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.

Newman, M. E. J. (2001d). Who is the best connected scientist? a study of scientific coauthorship networks. *Physical Review E*, 64:016132.

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*, pages 1199–1235.

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

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.

West, D. B. (2001). *Introduction to Graph Theory.* Prentice-Hall, 2nd edition.

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

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

Observe that $e_{is} = g_{is}\widetilde{e}_{is}$, where $\widetilde{e}_{is}$ denotes the "latent" effort level. The first order condition of maximizing utility (A.1) with respect to $\widetilde{e}_{is}$ gives

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

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

$$\mathbf{e} = (1+d)\mathbf{G}(\delta\otimes\alpha) + \lambda(1+d)\mathbf{W}\mathbf{e} - \phi\mathbf{M}\mathbf{e}.$$

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

B. Data Appendix

We use the following variables, retrieved in July 2017:

- 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.[32]
  4. Current RePEc ranking for the main affiliation fo the author.
  5. Year of the first publication recorded in the RePEc profile (article or paper).
  6. Year of completion of terminal degree, as listed in the RePEc Genealogy.
  7. Number of registered coauthors during career.
  8. Dummy for editor of journal.
  9. Dummy for NBER or CEPR affiliation.
  10. Dummy for terminal degree from an Ivy League institution.

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

11. Dummy for terminal degree obtained in the United States.
12. Dummy for main affiliation in the United States.
13. Gender as determined by a likelihood table using the first and possibly middle name. Uncertain matches were almost all resolved through internet search.
14. Ethnicity.
15. Closeness centrality measure.
16. Betweenness centrality measure.
17. Number of NEP fields in which author's work has been cited, to measure breadth of citations.
18. Inverse participation ratio for NEP fields of publications.
19. Fields of work, as determined by the NEP fields for which their working papers were selected for email dissemination.

- 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.
  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 last 12 months, as provided by LogEc.
  11. The number of authors.
  12. The average number of works across authors of this paper.
  13. Same, weighted by simple impact factors.
  14. Same, weighted by recursive impact factors.

15. The average number of citations across authors of this paper.
16. Same, weighted by simple impact factors.
17. Same, weighted by recursive impact factors.
18. The average number of citations across authors of this paper, each citation also divided by the number of authors of the cited paper.
19. Same, weighted by simple impact factors.
20. Same, weighted by recursive impact factors.
21. Number of references in the paper that have been matched with other items in RePEc.
22. Same, weighted by simple impact factors.
23. Same, weighted by recursive impact factors.
24. Year of publication in a journal.
25. Dummy if at least one author is editor.
26. Dummy if authors have main affiliations in different countries.

C. Heuristic Explanation of the Estimation Bias

In this appendix we provide a heuristic explanation on why the coefficients λ and ϕ estimated from Model (1) of Table 3 would be downward biased. First note that it is not straightforward to exploit data variations to gauge the direction of estimation bias because the variations of paper qualities, paper authorships, and author characteristics have been distorted nonlinearly in the equilibrium effort of Eq. (5). Alternatively, we consider a comparison between original and counterfactual scenarios. In the original scenario, we use the estimates of λ, ϕ, and β from Model (1) and Model (2), respectively, to predict papers' outputs. Intuitively, these predicted values are the best approximation of the real values that each model can offer. In the counterfactorual scenario, we take the estimated authors' abilities from Model (1), but manipulate the predicted research efforts and outputs by using the "higher" values of λ and ϕ from Model (2). The goal is to show that these manipulated outputs will be further apart from the real ones.

In Figure C.1 we contrast the author abilities obtained from Model (1) and from Model (2). Compared to Model (2), Model (1) overestimates the average value, but underestimates the dispersion. We believe that this is due to the fact that Model (1) omits the individual latent variables. Followed by this pattern, the computed equilibrium efforts from Model (1) also show a higher average but a smaller dispersion compared to Model (2), as shown in Figure C.2 (a). In addition, if we manipulate the computation of efforts from Model (1) by using the higher values of β and ρ from Model (2), Figure C.2 (b) shows that the efforts from Model (1) become even more concentrated and deviate further away from the result of Model (2). In Figure C.3 we continue to contrast the predicted paper outputs from Model (1) and Model (2). In panel (a), we can see that Model (2) provides better predictions than Model (1), despite that both Models fail to match a large amount of real values which cluster near zero. When we again manipulate the predicted paper outputs from Model (1) by using higher values of λ and ϕ from Model (2), the predictions of Model (1) become more concentrated at one middle range, which make the whole distribution further deviate from the true one. From these comparisons, we can conclude that due to omitting individual latent variables, Model (1) needs to underestimate the values of λ and ϕ in order to maintain a better goodness of fit to the real data.

D. Simulation Study

To show the proposed Bayesian MCMC estimation approach in Section 4.4 can effectively recover the true parameters from the model of Eqs. (14) and (15), we conduct a Monte Carlo

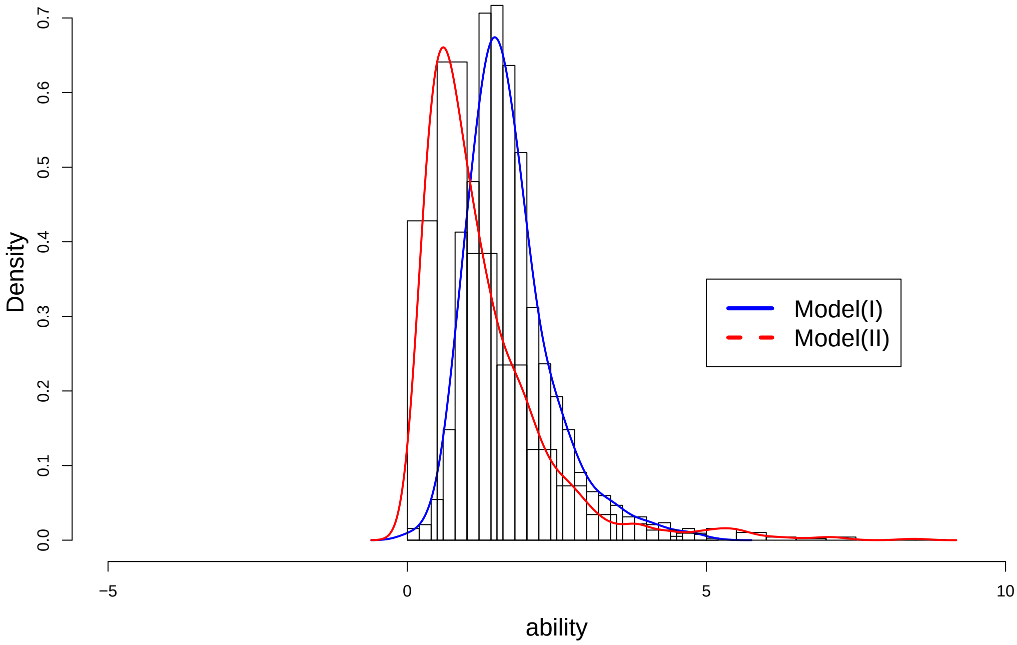
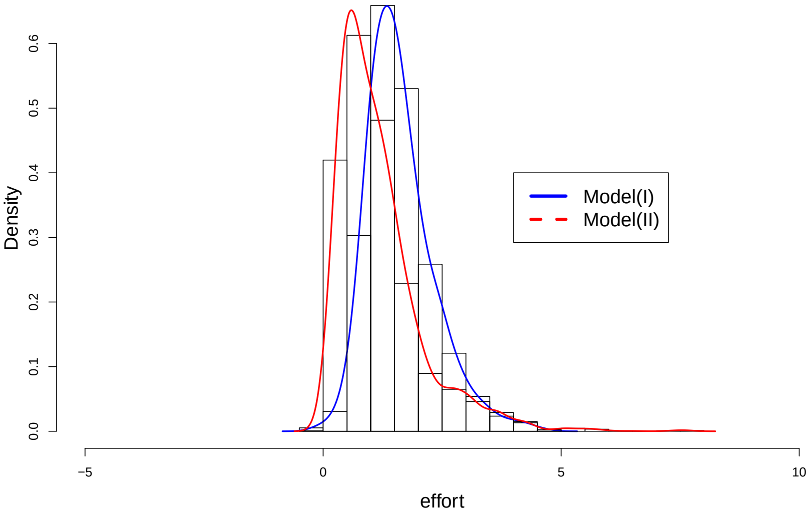


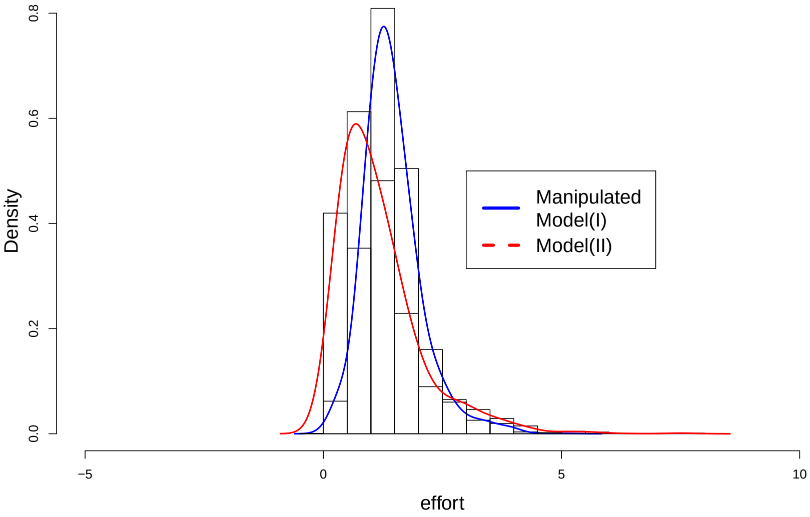
Figure C.1: Distributions of author abilities computed from the models of exogenous and endogenous project matching

simulation study to examine the bias and standard deviation from estimation results. The simulation consists of 100 repetitions. In each repetition, we first simulate dyadic binary exogenous variables $\mathbf{z_{is}}$ 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 $\mathbf{z_{is}} = 1$; otherwise, we set $\mathbf{z_{is}} = 0$. We simulate individual exogenous variables $\mathbf{x}$, author latent variables $\boldsymbol{\mu}$, and project latent variables κ from standard normal distributions. Then we generate the artificial project output $\mathbf{y}$ and participation $\mathbf{G}$ based on the data generating process (DGP) in Eqs. (14) and (15). We estimate two models, one is the full model (i.e., the DGP model) where both project output and project participation are endogenous and the other is just the project output equation by treating the collaboration 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 shown in Table D.9. We report the bias and the standard 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 is the same problem that we can reproduce from our empirical findings, which strengthens our argument that when omitting individual latent variables, the variation on authors' abilities will be underestimated and it 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 of small finite sample biases. However, these finite sample biases fade off when the sample size increases, which tells us that the proposed estimation algorithm has the desired finite sample performance.

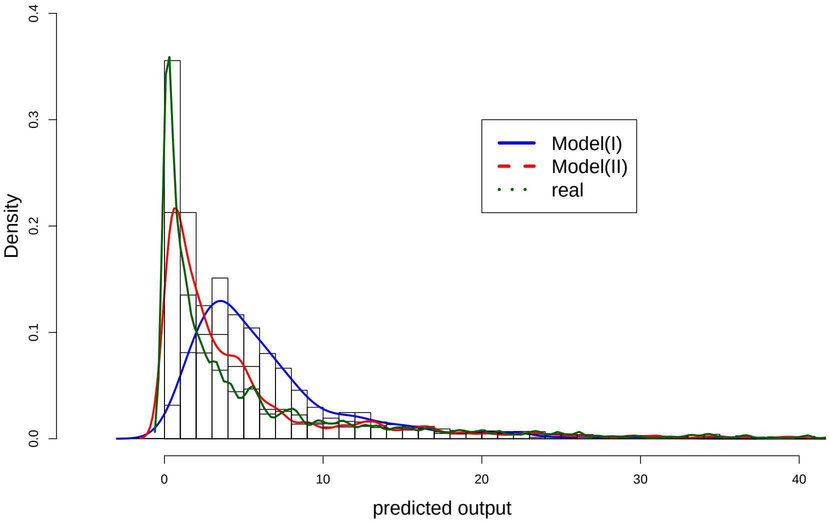


(a) original

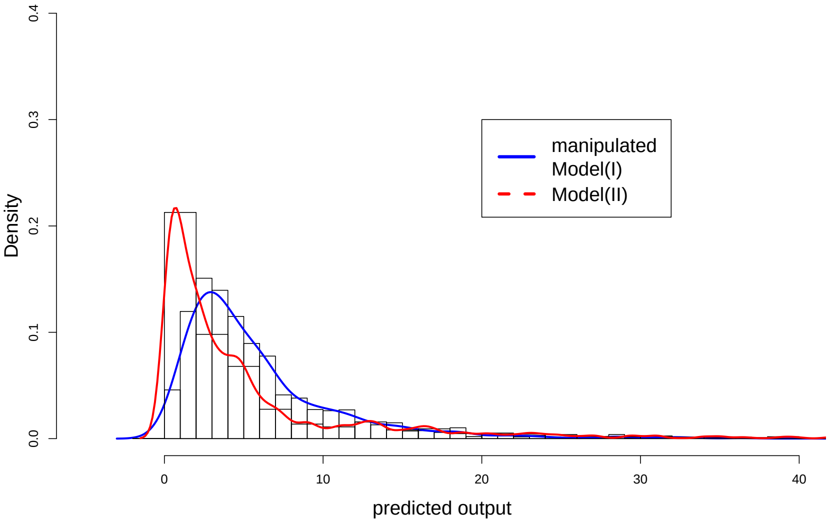


(b) manipulated

Figure C.2: Distributions of efforts computed from the models of exogenous and endogenous project matching



(a) original



(b) manipulated

Figure C.3: Distributions of paper qualities computed from the models of exogenous and endogenous project matching

Table D.9: Simulation results.

		n=200, p=250				n=300, p=350			
		Exogenous		Endogenous		Exogenous		Endogenous	
	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

E. Goodness-of-Fit Statistics

The matching model outlined in Section 4.3 attempts to uncover a channel in which authors choose projects 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 3. 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 artifical coauthor networks, we follow the iteration approach of Snijders (2002). In this approach, the simulated bipartite collaboration network G at different iterations t, $G^{(1)}, G^{(2)}, \cdots, G^{(t)}$, form a Markov chain and the transition probability of the Markov chain is given by

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

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

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

This M-H sampling procedure satisfies the detailed balance condition so that after convergence we can regard the realized 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 G, we do the projection to obtain the coauthor network adjacency matrix.

The examination results are shown in Figure E.1. We present the distribution of statistics for the observed network by solid curves, distributions for simulated networks by box plots and the 5^{th} and 95^{th} percentiles by dotted lines. 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 simulate the unobserved network generating process.

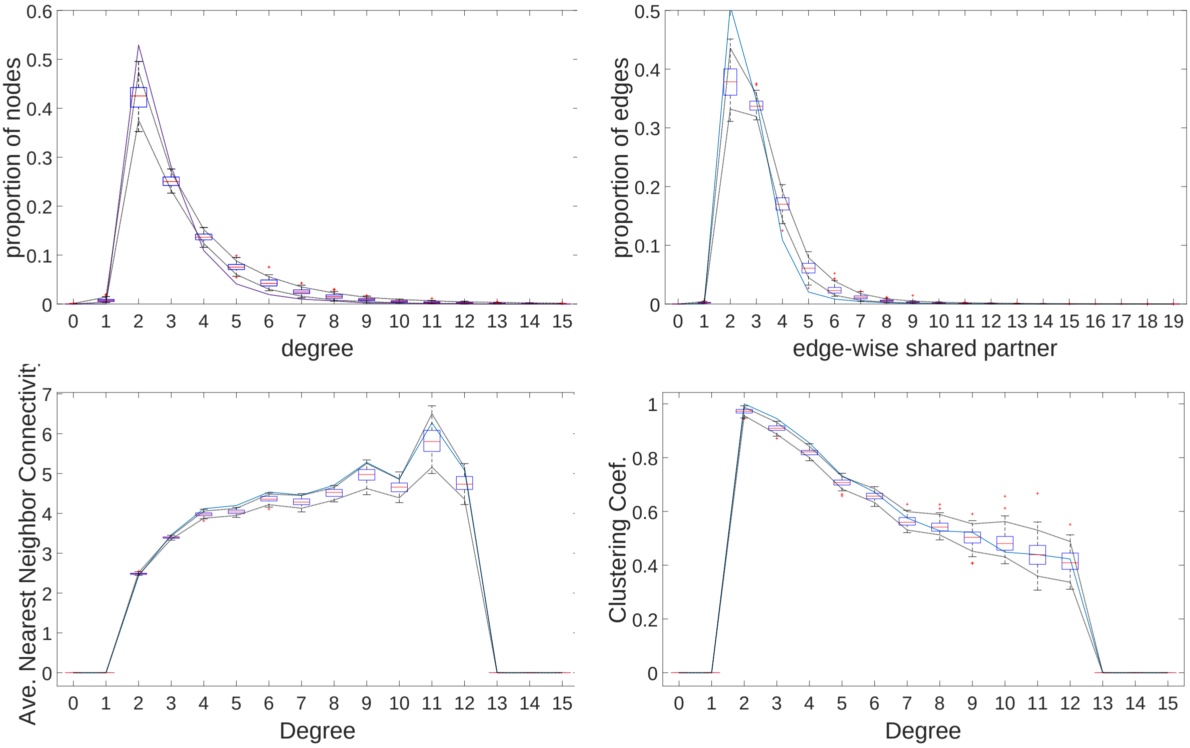


Figure E.1: Goodness-of-fit statistics for the coauthorship network.

F. Estimation Results for other Sample Periods

Table F.10: Summary statistics for the 2007-2009 sample.

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive Impact Factor	0.0000	91.2845	10.9982	17.8867	3601
number of authors (in each paper)	1	4	1.8298	0.7057	3601
Authors					
Log life-time citations	0	10.5394	5.7806	1.6097	1724
Decades after Ph.D. graduation	-0.8	5.30000	0.9959	0.9410	1724
Female	0	1	0.1340	0.3407	1724
NBER connection	0	1	0.1259	0.3318	1724
Ivy League connection	0	1	0.1618	0.3684	1724
Editor	0	1	0.0574	0.2327	1724
number of papers (for each author)	1	49	3.8219	3.8321	1724

Note: We drop authors who did not coauthor with any others during the sample period. We also drop papers without any citations when extracting from the RePEc data base.

Table F.11: Summary statistics for the 2013-2015 sample.

	Min	Max	Mean	S.D.	Sample size
Papers					
Citation recursive Impact Factor	0.0000	43.1194	3.1582	1.8687	1941
number of authors (in each paper)	1	5	1.9619	0.6891	1941
Authors					
Log life-time citations	0	10.5394	5.2776	1.8687	1301
Decades after Ph.D. graduation	-0.2	10.2000	1.2784	1.0005	1301
Female	0	1	0.1253	0.3312	1301
NBER connection	0	1	0.1238	0.3294	1301
Ivy League connection	0	1	0.1460	0.3533	1301
Editor	0	1	0.0507	0.2195	1301
number of papers (for each author)	1	52	2.9270	3.1572	1301

Note: We drop authors who did not coauthor with any others during the sample period. We also drop papers without any citations when extracting from the RePEc data base.

Table F.12: Estimation results for other sample periods.

	2007-2009		2013-2015	
	Model (1)	Model (2)	Model (1)	Model (2)
Output				
λ	-0.0692***	0.0509***	-0.0466	0.0555***
	(0.0228)	(0.0148)	(0.0443)	(0.0190)
ϕ	-0.0073	0.0251***	0.0015	0.1069***
	(0.0042)	(0.0061)	(0.0078)	(0.0134)
Constant	-2.2244***	-3.6015***	-0.9490***	-1.8399***
	(0.4534)	(0.3314)	(0.2856)	(0.1468)
Log life-time citations	0.7910***	0.7936***	0.4731***	0.5237***
	(0.0638)	(0.0554)	(0.0542)	(0.0393)
Decades after graduation	1.8056***	-0.1793	-0.4813	-1.8082***
	(1.0958)	(0.9251)	(0.4336)	(0.2644)
(Decades after graduation)^2	-3.9824***	-1.3618	-0.5675	1.5502***
	(1.3521)	(1.2348)	(0.6057)	(0.3889)
(Decades after graduation)^3	2.1675***	1.1673	0.4379	-0.4975***
	(0.7336)	(0.6834)	(0.3268)	(0.2185)
(Decades after graduation)^4	-0.4764***	-0.3052**	-0.1015	0.0586
	(0.1734)	(0.1616)	(0.0673)	(0.0450)
(Decades after graduation)^5	0.0362**	0.0244*	0.0064	-0.0022
	(0.0146)	(0.0136)	(0.0041)	(0.0027)
Female	0.3994**	0.1739	-0.2788	-0.2233**
	(0.2140)	(0.1828)	(0.2654)	(0.1212)
NBER connection	0.4219***	1.2190***	0.1163	0.0151
	(0.1430)	(0.1241)	(0.1254)	(0.1104)
Ivy League connection	0.8790***	0.7239***	0.1295	-0.2133**
	(0.1322)	(0.1171)	(0.1156)	(0.0943)
Editor	-0.7919***	-0.2316	-0.3160	-0.6087***
	(0.2687)	(0.2395)	(0.2160)	(0.1617)
ζ	–	6.2949***	–	3.4989***
		(0.3209)		(0.1559)
η	–	0.0352	–	-0.6512
		(0.9206)		(0.4871)
σ_v^2	235.4862***	152.8730***	21.6082***	11.0064***
	(5.6207)	(3.8297)	(0.7334)	(0.3974)
Matching				
Constant	–	-9.8859***	–	-9.0102***
		(0.1301)		(0.1351)
Same NEP	–	0.4036***	–	1.2067***
		(0.0510)		(0.0664)
Ethnicity	–	0.1559**	–	0.6788***
		(0.0853)		(0.1083)
Affiliation	–	1.8167***	–	4.6835***
		(0.1713)		(0.2298)
Female	–	-0.3007***	–	-0.0020
		(0.1018)		(0.1236)
Advisor-advisee	–	1.3886***	–	5.1268***
		(0.1119)		(0.1863)
Past coauthors		7.8126***		5.1480***
		(0.1051)		(0.1236)
Share common co-authors	–	2.0309***	–	2.2454***
		(0.3542)		(0.2426)
Author effect	–	3.0220***	–	3.7976***
		(0.1949)		(0.2097)
Project effect	–	0.0533	–	0.2687
		(0.2951)		(0.2659)
Sample size	3601		1941	

Note: Model (1): assume exogenous matching between authors and papers. Model (2): assume endogenous matching by Equation (13).