國立臺灣大學經濟學系
Department of Economics, National Taiwan University

NTU WORKING PAPER SERIES

Collaboration in Bipartite Networks

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

NTU Working Paper No. 2202

DEPARTMENT OF ECONOMICS
NATIONAL TAIWAN UNIVERSITY
https://econ.ntu.edu.tw/

Collaboration in Bipartite Networks*

Chih-Sheng Hsieh† Michael D. König‡ Xiaodong Liu§

Christian Zimmermann¶

February 16, 2022

*This paper was previously circulated under the title "Superstar economists: coauthorship networks and research output". We thank Pietro Biroli, Francis Bloch, Pierre-Philippe Combes, Lorenzo Ductor, Kenan Huremovic, Matt Jackson, Carlos Martins-Filho, Adam McCloskey, David Miller, Ralph Ossa, Seth Richards-Shubik, Hannes Schwandt, Marco Van Der Leij, Bauke Visser, Joachim Voth, Fabrizio Zilibotti and seminar participants at Baptist University of Hong Kong, Chinese University of Hong Kong, Jean Monnet University in St-Étienne, Laval University, Ludwig Maximilian University of Munich, National Taiwan University, Ohio State University, Tinbergen Institute, University of Zurich, Xiamen University, the Econometric Society Meeting in Barcelona, the NSF Conference on Network Science and Economics at Washington University, and the Workshop on the Economics of Scientific Research at Erasmus University Rotterdam for their helpful comments. Moreover, we thank Adrian Etter and Marc Biedermann for excellent research assistance. Michael D. König acknowledges financial support from the Swiss National Science Foundation through research grant PZ00P1_154957 /1. The views expressed are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

†Department of Economics, National Taiwan University, Taipei 10617, Taiwan. Email: cshsieh@ntu.edu.tw.

‡Centre for Economic Policy Research (CEPR), London, United Kingdom. ETH Zurich, Swiss Economic Institute (KOF), Zurich, Switzerland. Department of Spatial Economics, VU Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Email: m.d.konig@vu.nl.

§Department of Economics, University of Colorado Boulder, Boulder, Colorado 80309-0256, United States. Email: xiaodong.liu@colorado.edu.

¶Department of Economic Research, Federal Reserve Bank of St. Louis, St. Louis MO 63166-0442, United States. Email: zimmermann@stlouisfed.org.

Abstract

This paper studies the impact of collaboration on research output. First, we build a micro-founded model for scientific knowledge production, where collaboration between researchers is represented by a bipartite network. The Nash equilibrium of the game incorporates both the complementarity effect between collaborating researchers and the substitutability effect between concurrent projects of the same researcher. Next, we propose a Bayesian MCMC procedure to estimate the structural parameters, taking into account the endogenous participation of researchers in projects. Finally, we illustrate the empirical relevance of the model by analyzing the coauthorship network of economists registered in the RePEc Author Service. The estimated complementarity and substitutability effects are both positive and significant when the endogenous matching between researchers and projects is controlled for, and are downward biased otherwise. To show the importance of correctly estimating the structural model in policy evaluation, we conduct a counterfactual analysis of research incentives. We find that the effectiveness of research incentives tends to be understated when the complementarity effect is ignored and overstated when the substitutability effect is ignored.

Keywords: bipartite networks, coauthorship networks, research collaboration, spillovers, economics of science.

JEL: C31, C72, D85, L14

1 Introduction

Collaboration between researchers has played a significantly important role in economics in recent decades. In 2014, multi-authored papers accounted for 75% of all articles published in economics (Kuld & O'Hagan 2018).[1] Through a complex network of collaborations, researchers generate spillovers not only to their coauthors but also to other researchers indirectly connected to them. The aim of this paper is to develop a structural model that helps us to understand how collaboration affects research output.

First, we build a micro-founded model for scientific knowledge production. The collaboration between researchers is characterized by a bipartite network with two types of nodes: researchers and research projects. The effort that a researcher spends in a project is represented by an edge in the bipartite network, and collaborating researchers are connected through the projects they work on together. We characterize the equilibrium of the game where researchers choose efforts in multiple and possibly overlapping projects to maximize utility. The equilibrium takes into account both the complementarity effect between collaborating researchers and the substitutability effect between concurrent projects of the same researcher.

Next, we propose an estimation procedure to recover the structural parameters of the model. There are three main challenges in estimating this model. First, the effort level of a researcher in the production function is unobservable. To overcome this problem, we replace the unobserved effort level in the production function with the equilibrium effort level derived from the theoretical model. Second, the matching between researchers and projects is likely to be endogenous. Estimating the production function without taking into account this potential endogeneity may incur a selection bias. To remediate the issue, we introduce a participation function to model the endogenous selection of researchers into projects, allowing for both researcher and project unobserved heterogeneity.[2] The resulting

[1] Additional evidence can be found in Ductor (2015).

[2] As pointed out in Bonhomme (2020), a key feature of bipartite networks is two-sided heterogeneity.

likelihood function involves high-dimensional integrals. This leads to the third challenge of the estimation, i.e., it is computationally cumbersome to apply a frequentist maximum likelihood method, even when resorting to a simulation approach. To bypass this difficulty, we adopt a Bayesian Markov Chain Monte Carlo (MCMC) approach to jointly estimate the production and participation functions.

Finally, we bring our model to the data by analyzing the coauthorship network of economists registered in the Research Papers in Economics (RePEc) Author Service. We find that the estimated complementarity and substitutability effects are both statistically significant with the expected signs. The estimates are downward biased when the endogenous matching between researchers and projects is ignored. The direction of the bias is compatible with the intuition and consistent with the Monte Carlo simulation results. We also conduct a series of robustness checks to explore the sensitivity of our results to alternative specifications and samples. To illustrate the importance of correctly estimating the structural model in policy analysis, we carry out a counterfactual study on the impact of research incentives on research output. We find that the effectiveness of research incentives tends to be understated when the complementarity effect is ignored and overstated when the substitutability effect is ignored.

There exists a growing literature, both empirical and theoretical, on the formation and impact of scientific collaboration networks. On the empirical side, the structural features of scientific collaboration networks have been analyzed in Newman (2001*c*,*a*,*b*, 2004*a*,*b*) and Goyal et al. (2006). Fafchamps et al. (2010) examine predictors for the establishment of scientific collaborations. Azoulay et al. (2010) estimate the negative externality induced by the premature and sudden death of active "superstar" scientists on their coauthors. Ductor et al. (2014), Ductor (2015), Ductor et al. (2021), and Lindenlaub & Prummer (2021) study how intellectual collaboration affects the research output of individual authors. Anderson & Richards-Shubik (2021) use a strategic network formation model to study how researchers choose their collaborators and the projects they work on. Bonhomme (2021) proposes an

econometric framework to identify individual contributions to the output of their teams, by tracking researchers who work in different teams over time. In this paper, we provide a structural model characterizing how researchers allocate their effort across different projects taking into account the complementarity of collaboration and the substitutability of their own research effort in different projects. It is important to quantify both the complementarity and substitutability effects as they provide two different channels of connection in the bipartite network. As we demonstrate in the counterfactual study, correctly estimating their magnitudes is essential for policy evaluation.

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 setup considered in these papers by allowing for investments into multilateral relationships. Moreover, in a related paper Bimpikis et al. (2019) 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. While the products sold by competing firms to the same market are substitutes in Bimpikis et al. (2019), the efforts spent by collaborating agents in the same project are strategic complements in our model.

The rest of the paper is organized as follows. Section 2 introduces the theoretical model and characterizes the equilibrium. Section 3 presents the econometric methodology. The empirical implications of the model are discussed in Section 4, where Section 4.1 describes the data used in the empirical study, Section 4.2 gives the main estimation results, Section 4.3 reports the estimated marginal effects, Section 4.4 provides robustness analysis, and Section 4.5 conducts a counterfactual study on research incentives. Section 5 briefly concludes. The proofs, technical details, and additional robustness checks can be found in the online appendix.

2 Theoretical Model

2.1 Bipartite Network, Production Function, and Utility

Consider a *bipartite* network given by $\mathcal{G} = (\mathcal{N}, \mathcal{P}, \mathcal{E})$, where $\mathcal{N} = \{1, \ldots, n\}$ denotes the set of agents, $\mathcal{P} = \{1, \ldots, p\}$ denotes the set of projects, and $\mathcal{E}$ denotes the set of edges connecting agents and projects. In our model, an edge $e_{is} \in \mathcal{E}$ is the (non-negative) effort that agent i spends in project s. Let $\mathcal{N}_s$ denote the set of agents working on project s and $\mathcal{P}_i$ denote the set of projects agent i participates in. Let $|\cdot|$ denote the cardinality of a set.

The *production function* for project $s \in \mathcal{P}$ is given by

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

where $y_s(\mathcal{G})$ (or simply y_s) is the output of project s, α_i represents individual heterogeneity in productivity, $g_{ij} \in [0, 1]$ measures the degree of compatibility between collaborating agents, and ϵ_s is a random shock. If λ is positive, then the marginal product of agent i's effort in a project increases with the efforts of other agents in that project. Hence, the coefficient λ captures the complementarity effect.

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

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

The utility function has a payoff/cost structure. The payoff is the weighted total output of the projects that agent i participates in, with the weights given by $\delta_s \in (0, 1]$.[3] The cost is quadratic in efforts, with the coefficient ϕ measuring the degree of substitutability of an agent's efforts in different projects. If ϕ is positive, then the marginal cost of agent i's effort

[3]For example, if $\delta_s = 1/|\mathcal{N}_s|$, then the individual payoff is discounted by the number of agents participating in project s (cf. Kandel & Lazear 1992, Jackson & Wolinsky 1996, Hollis 2001).

in a project increases with the effort agent i spends on other projects. This quadratic cost specification helps to capture the fact that the time or resource of a researcher is limited.[4] It includes the convex separable cost specification as a special case with $\phi = 0$. The quadratic cost specification is very common in the literature (Singh & Vives 1984). A theoretical model with a similar cost specification but allowing for only two activities is studied in Belhaj & Deroïan (2014) and an empirical analysis is provided in Liu (2014) and Cohen-Cole et al. (2018). In addition, a convex separable cost specification can be found in the model studied in Adams (2006).

2.2 Game and Equilibrium

Prior to the effort-allocation game, we assume agents randomly meet with each other and come up with research ideas/projects.[5] This stochastic meeting process takes into account assortativity and homophily, reflecting the fact that "similarity breeds connection" (McPherson et al. 2001, Currarini et al. 2009). The outcomes of the meeting process are characterized by indicator variables d_{is}, such that $d_{is} = 1$ if agent i is in project s and $d_{is} = 0$ otherwise. Given $\{d_{is}\}$, agents strategically allocate research efforts $e_{is} \geq 0$ to the projects that they participate in to maximize utility in the effort-allocation game.

The following proposition provides an equilibrium characterization of the agents' effort portfolio $e = (e_1', \cdots, e_p')'$, with $e_s = (e_{1s}, \cdots, e_{ns})'$ for $s = 1, \cdots, p$. Let

$$W = D(\text{diag}_{s=1}^{p}\{\delta_s\} \otimes G)D, \qquad \text{and} \qquad M = D(J_p \otimes I_n)D, \tag{3}$$

where $\otimes$ denotes the Kronecker product, D is an np-dimensional diagonal matrix given by $D = \text{diag}_{s=1}^{p}\{\text{diag}_{i=1}^{n}\{d_{is}\}\}$, G is an $n \times n$ zero-diagonal matrix with the (i,j)th $(i \neq j)$ element being g_{ij}, and J_p is an $p \times p$ zero-diagonal matrix with off-diagonal elements equal

[4]For example, Ductor (2015) finds evidence for a congestion externality proxied by the average number of coauthors' papers that has a negative effect on individual academic productivity.

[5]It is possible for the same group of researchers to come up with multiple research ideas/projects.

to one. Let $\rho_{\max}(\cdot)$ denote the spectral radius of a square matrix.

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). Let $L := L(\lambda, \phi) = \lambda W - \phi M$. Given $\{d_{is}\}$, if*

$$\rho_{\max}(L) < 1, \tag{4}$$

then the Nash equilibrium effort portfolio is given by

$$e^* = (I_{np} - L)^{-1} D(\delta \otimes \alpha), \tag{5}$$

where $\delta = (\delta_1, \cdots, \delta_p)'$ and $\alpha = (\alpha_1, \cdots, \alpha_n)'$.

The matrix $L = \lambda W - \phi M$ represents a weight matrix of the *line graph* $\mathcal{L}(\mathcal{G})$ for the bipartite network $\mathcal{G}$.[6] In the line graph $\mathcal{L}(\mathcal{G})$, each node represents the effort an agent invests into a project. The links between nodes with the same project are represented by the nonzero entries of W while the links between nodes with the same agent are represented by the nonzero entries of M. The matrix L is a weighted sum of the matrices W and M, with the weights being the complementarity effect (λ) and the substitutability effect (ϕ) respectively. The formulation of L highlights the importance of both effects (i.e., λ and ϕ) in the bipartite network. The condition in Equation (4) plays a similar role as the one in Theorem 1 of Ballester et al. (2006), which limits the rate spillovers decay across the bipartite network.

2.3 An Illustrating Example

We illustrate the equilibrium characterization of Proposition 1 with an example corresponding to the bipartite network $\mathcal{G}$ in Figure 1. In this bipartite network, there are 3 agents and

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

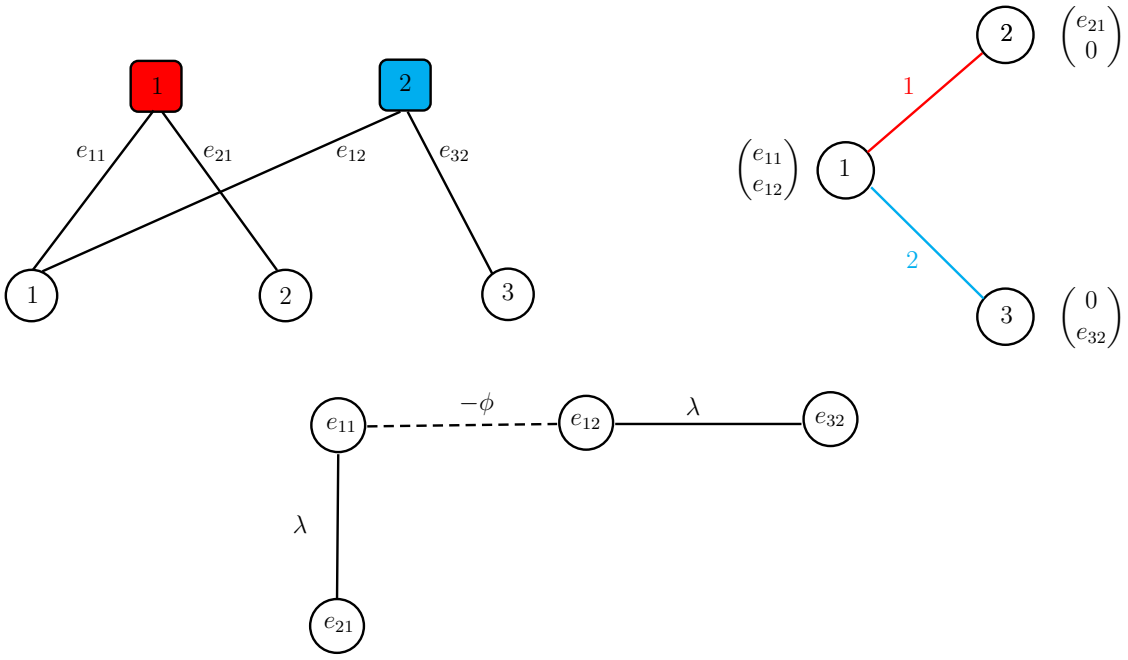

Figure 1: Top left panel: the bipartite network $\mathcal{G}$ of agents and projects analyzed in Section 2.3, where circles represent agents and squares represent projects. Top right panel: the projection of the bipartite network $\mathcal{G}$ on the set of agents. The effort levels of the agents for each project they are involved in are indicated next to the nodes. Bottom panel: the line graph $\mathcal{L}(\mathcal{G})$ associated with the bipartite network $\mathcal{G}$, in which each node represents the effort an agent invests into a project. Solid lines connect nodes with the same project while dashed lines connect nodes with the same agent.

2 projects, where agents 1 and 2 are collaborating in the first project and agents 1 and 3 are collaborating in the second project. For expositional purposes, let $g_{ij} = 1$ for all $i \neq j$ and $\delta_s = 1$ for all s.

Line Graph. The line graph $\mathcal{L}(\mathcal{G})$ of this bipartite network is depicted in the bottom panel of Figure 1. In the line graph, each node represents the effort an agent invests into a project. Solid lines connect nodes with the same project while dashed lines connect nodes with the same agent. Following Equation (3),

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

The nonzero entries of the matrices W and M correspond to, respectively, the solid lines and the dashed lines in the line graph. The matrix L is a weighted sum of the matrices W and M, given by

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

The $(1,2)$th and $(2,1)$th elements of the matrix L 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 $-\phi$. It is worth pointing out that, in the absence of the substitutability effect (i.e., $\phi = 0$), the line graph would be split into two independent sub-graphs with each one corresponding to the collaborators' efforts in a single project. Therefore, the substitutability effect provides a channel to capture the interdependence of efforts in different projects.

Equilibrium. In this example, the sufficient condition (4) for the existence of a unique equilibrium holds if $|\phi| < 1 - \lambda^2$. This condition requires both the complementarity effect λ and the substitutability effect ϕ to be less than one. Note that this condition reduces to $|\phi| < 1$ if $\lambda = 0$ and $|\lambda| < 1$ if $\phi = 0$. From Equation (5), the equilibrium effort portfolio is

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

Marginal Effects of α_i. As $|\phi| < 1 - \lambda^2$,

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

and, if the complementarity effect is positive (i.e., $\lambda > 0$),

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

which suggest that more productive agents raise not only their own effort levels but also the effort levels of their collaborators due to the complementarity effect. On the other hand, if the substitutability effect is also positive (i.e., $\phi > 0$),

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

which suggest that more productive agents induce lower effort levels spent by agents on other projects. It is worth noting that, without the substitutability effect (i.e., $\phi = 0$), agent i's productivity would have no effect on other agents' effort levels on a project that agent i is not involved in. This spotlights the important role of the substitutability effect in the bipartite network. An illustration can be seen in Figure 2.

Marginal Effects of λ. The partial derivative of the equilibrium effort of agent 1 in project 1 with respect to the complementarity 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 substitutability effect ϕ is also large, agent 1 may spend even less effort in the project with agent 2, leading to a negative $\partial e_{11}^*/\partial\lambda$. An illustration can be seen in Figure 3.

Marginal Effects of ϕ. The partial derivatives of the equilibrium efforts of agent 1 in projects 1 and 2 with respect to the substitutability coefficient ϕ are given by

$$\frac{\partial e_{11}^*}{\partial \phi} = -\frac{1}{2}\left[\frac{\lambda(\alpha_3 - \alpha_2)}{(1-\lambda^2-\phi)^2} + \frac{2\alpha_1 + \lambda(\alpha_2+\alpha_3)}{(1-\lambda^2+\phi)^2}\right],$$
$$\frac{\partial e_{12}^*}{\partial \phi} = \frac{1}{2}\left[\frac{\lambda(\alpha_3 - \alpha_2)}{(1-\lambda^2-\phi)^2} - \frac{2\alpha_1 + \lambda(\alpha_2+\alpha_3)}{(1-\lambda^2+\phi)^2}\right].$$

Suppose $\alpha_3 > \alpha_2$. Then, $\partial e_{11}^*/\partial\phi$ is negative. That is, with increasing ϕ, agent 1 exerts lower effort in the project with a less productive collaborator. In contrast, $\partial e_{12}^*/\partial\phi$ can be positive or negative, depending on whether the first term is larger or smaller than the second term on the right hand side of the second equation. With $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$, and $\lambda = 0.2$, we can see in Figure 4 that, when the substitutability effect ϕ is small, both $\partial e_{11}^*/\partial\phi$ and $\partial e_{12}^*/\partial\phi$ are negative, and $\partial e_{11}^*/\partial\phi < \partial e_{12}^*/\partial\phi$. That is, increasing ϕ reduces efforts of agent 1 in both projects, and the effort reduction is more significant in the project with a less productive collaborator. When ϕ is larger, $\partial e_{12}^*/\partial\phi$ becomes positive while $\partial e_{11}^*/\partial\phi$ remains negative, indicating agent 1 reallocates effort to the project with a more productive collaborator as a result of the substitutability effect.

3 Estimation

Recall $d_{is} = \mathbb{1}(i \in \mathcal{N}_s)$, where $\mathbb{1}(\cdot)$ denotes an indicator function. Equation (1) can be rewritten as

$$y_s = \sum_{i\in\mathcal{N}} \alpha_i d_{is} e_{is} + \frac{\lambda}{2}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}\setminus\{i\}} g_{ij} d_{is} d_{js} e_{is} e_{js} + \epsilon_s, \tag{6}$$

where ϵ_s is i.i.d.$(0, \sigma_\epsilon^2)$. In the empirical model, we assume agent i's productivity is given by

$$\alpha_i = \exp(x_i'\beta), \tag{7}$$

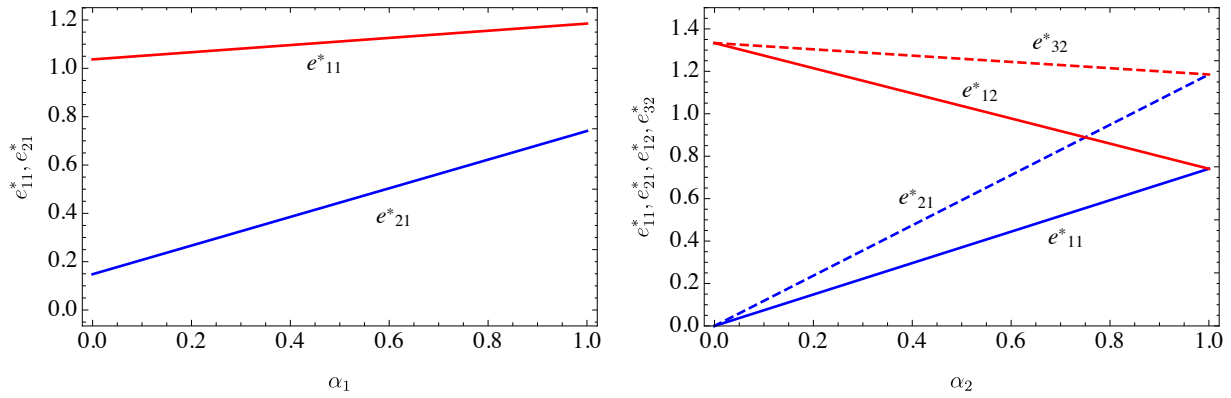

Figure 2: Left panel: equilibrium effort levels for agents 1 and 2 in project 1 for $\lambda = 0.25$, $\phi = 0.75$, $\alpha_2 = \alpha_3 = 1$, and varying values of α_1. (In this case, $e^*_{11} = e^*_{12}$ and $e^*_{21} = e^*_{32}$.) Right panel: equilibrium effort levels for agents 1, 2 and 3 in projects 1 and 2 for $\lambda = 0.25$, $\phi = 0.75$, $\alpha_1 = \alpha_3 = 1$, and varying values of α_2.

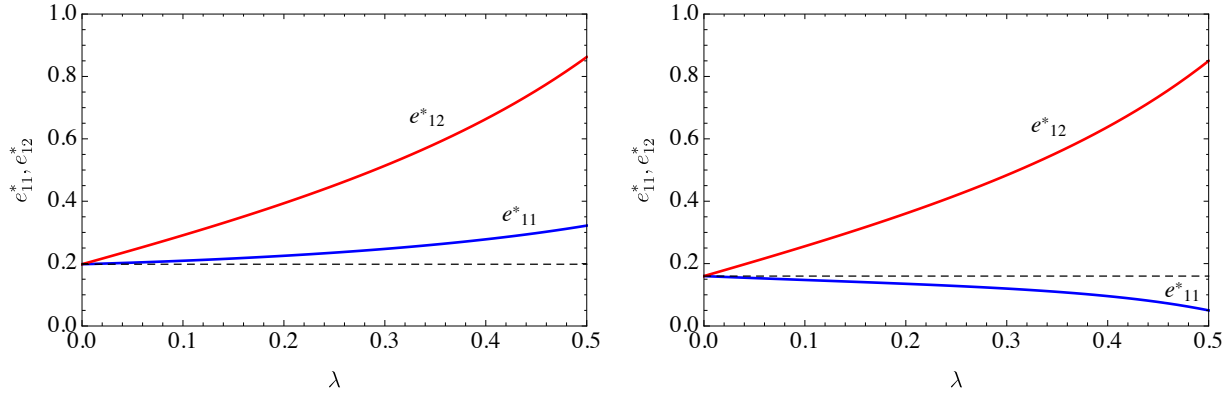

Figure 3: Equilibrium effort levels for agent 1 with $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$, $\phi = 0.05$ (left panel), and $\phi = 0.25$ (right panel), for varying values of λ. The dashed lines indicate the effort levels for $\lambda = 0$.

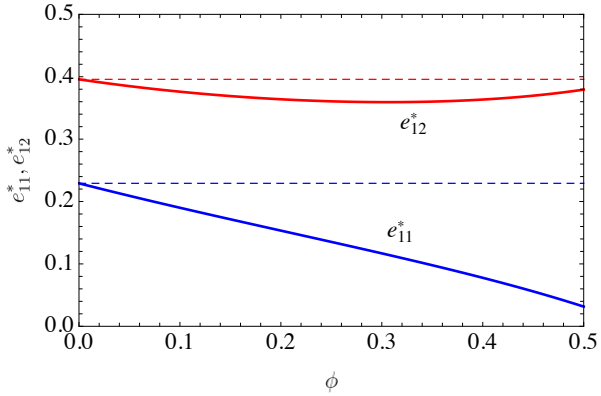

Figure 4: Equilibrium effort levels for agent 1 with $\alpha_1 = 0.2$, $\alpha_2 = 0.1$, $\alpha_3 = 0.9$, and $\lambda = 0.2$, for varying values of ϕ. The dashed lines indicate the effort levels for $\phi = 0$.

where x_i is a vector of observable individual attributes. Equation (7) is assumed to be an exponential function to guarantee that the productivity is positive.

There are three main challenges in estimating this model. First, the effort level e_{is} is usually unobservable to the econometrician. To overcome this problem, we replace e_{is} in Equation (6) with the equilibrium effort level e^*_{is} given by Equation (5) and estimate the equilibrium production function

$$y_s = \sum_{i \in \mathcal{N}} \alpha_i d_{is} e^*_{is} + \frac{\lambda}{2} \sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N} \setminus \{i\}} g_{ij} d_{is} d_{js} e^*_{is} e^*_{js} + \epsilon_s. \tag{8}$$

Equation (8) is highly nonlinear in the unknown parameters. Thus, it is difficult to derive easy-to-check sufficient conditions for identification as in Bramoullé et al. (2009). To get some intuition on what data variation identifies the complementarity parameter λ and the substitutability parameter ϕ respectively, we consider the exemplary networks in Figure 5. The first network has two agents collaborating on a joint project, the second network has one agent working on two projects, and the last network has two agents each working on a solo project. Suppose $\delta_s = 1$ and the productivities (α_i) of all the agents are identical. If $\lambda = 0$, then the (expected) equilibrium total outputs of the first and the last networks would be the same. Thus, λ can be identified from the output variation of these two networks. Similarly, if $\phi = 0$, then the (expected) equilibrium total outputs of the second and the last networks would be the same. Thus, ϕ can be identified from the output variation of these two networks. Therefore, in the real data, when the structure of the bipartite network is sufficiently rich, we should be able to identify both complementarity and substitutability effects.

Second, d_{is} is likely to be endogenous. For example, in a coauthorship network, high-ability researchers are more likely to work on different projects at the same time, and high-potential projects are usually harder to find and more challenging to work on. Furthermore, researchers tend to be sorted into projects based on their research interests and abilities.

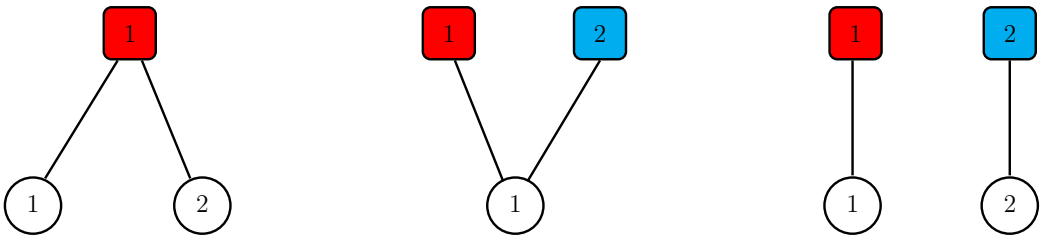

Figure 5: Exemplary networks with two agents or two projects, respectively.

Estimating Equation (8) without taking into account potential endogeneity of d_{is} may incur a selection bias. To control for the endogenous selection, we introduce a participation function allowing for both agent and project unobserved heterogeneity. More specifically, we assume

$$d_{is} = \mathbb{1}(z_{is}'\gamma + \xi\mu_i + \psi\eta_s + \kappa|\mu_i - \eta_s| + v_{is} > 0), \tag{9}$$

where z_{is} is a vector of observables measuring compatibility between agent i and project s,[7] μ_i is an i.i.d.$(0,1)$ agent-specific random component, η_s is an i.i.d.$(0,1)$ project-specific random component, and v_{is} is an i.i.d.$(0,1)$ error term independent of μ_i and η_s. As the number of observations d_{is} is much larger than the number of random components μ_i and η_s, it is reasonable to assume that μ_i and η_s can be identified from Equation (9). To allow the agent and project unobserved heterogeneity to also affect production, we assume agent i's productivity depends on the agent-specific random component μ_i so that Equation (7) becomes

$$\alpha_i = \exp(x_i'\beta + \zeta\mu_i),$$

and the error term in Equation (8) can be written as

$$\epsilon_s = \varsigma\eta_s + u_s,$$

[7]In the empirical illustration, z_{is} includes the Jaffe similarity between agent i's research fields and project s's fields as a measure of the compatibility between the agent and project. z_{is} also includes terms capturing the connections between agent i and other agents collaborating in project s in terms of affiliation, past coauthorship, etc. The idea is that for the agents that are in the same affiliation or coauthors in the past, they are more likely to meet and come up with new ideas/projects together. See Section 4.2 for more details.

where u_s is an i.i.d.$(0, \sigma_u^2)$ error term independent of η_s. This specification has the following implications. First, if $\zeta > 0$ and $\xi > 0$, then a researcher with higher ability (given by a higher μ_i) tends to participate in more projects. Second, if $\varsigma > 0$ and $\psi < 0$, then a project with higher potential (given by a higher η_s) has a higher threshold for researchers to participate in. Finally, if $\kappa < 0$, then agents are more likely to join projects that match their abilities, i.e., agents are sorted into projects based on homophily of unobserved characteristics.

Third, with the unobserved heterogeneity, the joint likelihood function of production and participation involves high-dimensional integrals and is computationally cumbersome to evaluate. To bypass this difficulty, we follow the Bayesian approach of Zeger & Karim (1991). Let $\theta_d = (\gamma', \xi, \psi, \kappa)'$ and $\theta_y = (\lambda, \phi, \beta', \zeta, \varsigma, \sigma_u^2)'$. Let $f(d|\mu, \eta, \theta_d)$ denote the conditional probability of $d = [d_{is}]$ given $\mu = (\mu_1, \cdots, \mu_n)'$ and $\eta = (\eta_1, \cdots, \eta_p)'$, and $f(y|d, \mu, \eta, \theta_y)$ denote the conditional density of $y = (y_1, \cdots, y_p)'$ given d, μ, and η. Then, μ, η, and $\theta = (\theta_y', \theta_d')'$ can be sampled from the joint posterior density

$$p(\mu, \eta, \theta|y, d) \propto f(y|d, \mu, \eta, \theta_y) f(d|\mu, \eta, \theta_d) \pi(\mu)\pi(\eta)\pi(\theta_y)\pi(\theta_d),$$

with the priors $\pi(\mu)$, $\pi(\eta)$, $\pi(\theta_y)$ and $\pi(\theta_d)$. The details of Bayesian estimation can be found in Appendix B.

4 Empirical Study: Coauthorship Networks

4.1 Data

The data used for this study make extensive use of the metadata assembled by the RePEc initiative and its various projects. RePEc assembles information about publications relevant to economics from over 2,000 publishers, including all major commercial publishers and university presses, policy institutions, and pre-prints (working papers) from academic

institutions.[8]

In addition, we make use of the data made available by various projects that build on these RePEc data and enhance it in various ways. First, we take the publication profiles of economists registered with the RePEc Author Service, which include what they have published and where they are affiliated.[9] Second, we extract information about their advisors, students, and alma mater, as recorded in the RePEc Genealogy project.[10] This academic genealogy data has been complemented with some of the data used in Colussi (2017).[11] Third, we use the New Economics Papers (NEP) project to identify the field-specific mailing lists through which the papers have been disseminated.[12] NEP has human editors who determine the field in which new working papers belong. We obtain 99 distinct NEP fields. Fourth, we use citations to the papers and articles as extracted by the CitEc project.[13] Finally, we use journal impact factors, as well as author and institution rankings from IDEAS.[14]

Compared with other data sources, RePEc has the advantage of linking these various datasets in a seamless way that is verified by the respective authors. Author identification is superior to any other dataset as homonyms are disambiguated by the authors themselves as they register and maintain their accounts. While not every author is registered, most are. Indeed, 90% of the top 1000 economists as measured by their publication records for the 1990-2000 period are registered.[15] We believe that the proportion is higher for the younger generation that is more familiar with social networks and online tools and thus more likely to register with online services.

In terms of publications, RePEc covers all important outlets and over 3,000 journals are listed, most of them with extensive coverage. References are extracted for about 30% of

[8] See `http://repec.org/` for a general description of RePEc.
[9] RePEc Author Service: `https://authors.repec.org/`
[10] RePEc Genealogy project: `https://genealogy.repec.org/`
[11] We would like to thank Tommaso Colussi for sharing the data with us.
[12] NEP project: `https://nep.repec.org/`
[13] CitEc project: `http://citec.repec.org/`
[14] IDEAS: `https://ideas.repec.org/top/`. For a detailed description of the factors and rankings, see Zimmermann (2013).
[15] `https://ideas.repec.org/coupe.html`

their articles (in addition to working papers) to compute citation counts and impact factors. The missing references principally come from publishers refusing to release them for reasons related to copyright protection. While the resulting gap is unfortunate, it is unlikely to result in a bias against particular authors, fields, or journals. The exception may be authors who are significantly cited in outlets outside of economics that may or may not be indexed in RePEc (note that several top management, statistics, and political science journals are also indexed).

To obtain a sample from RePEc that is appropriate for our analysis, we apply a series of filters as follows. First, we select papers that had a first pre-print version in 2010-2012. We choose 2010-2012 because it is old enough to give all authors a chance to have added the papers to their profiles and for the papers to have been eventually published in journals; but not too old for a good data coverage, as the coverage of RePEc becomes slimmer with older vintages. Furthermore, we require all authors of the papers to be registered with RePEc and all authors to have the RePEc Genealogy information on where they studied. We drop all duplicate or older versions of each paper from our sample. This gives us a sample of 6,673 papers written by 3,700 distinct authors for which we have complete data.

Next, as we use citations to measure research output, we drop 2,463 papers that do not have any citations up to November 2018 when the data is extracted from the RePEc database, as well as 658 authors who only work on these dropped papers without any citations. This reduces to the sample size to 4,210 papers and 3,042 authors.[16]

Finally, as we are interested in collaborations between researchers, we drop 621 authors who wrote only single-authored papers in the sample period. This results in a final sample of 3,589 papers written by 2,421 distinct authors.[17]

[16]In Appendix E, we conduct a robustness check by estimating the empirical model with a sample including the 2,463 papers without any citations. The main result is qualitatively unchanged.

[17]In Appendix E, we conduct a robustness check by estimating the empirical model with a sample including the 621 authors who wrote only single-authored papers in the sample period. The main result is qualitatively unchanged.

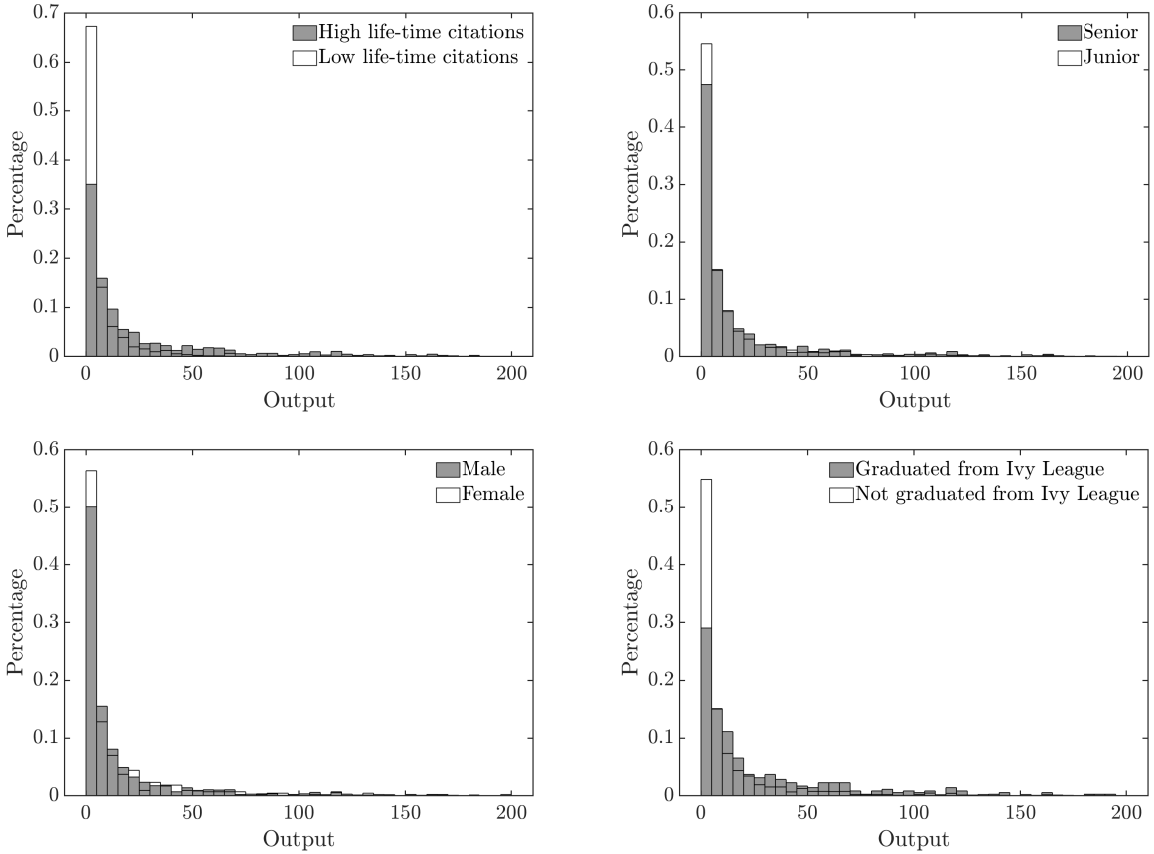

Figure 6: Distributions of research output for different types of researchers. In the top left figure, the cutoff point between high life-time citations and low life-time citations is the median number of lifetime citations, which is 300. In the top right figure, the cutoff point between senior and junior researchers is the median number of years after graduation, which is 9.

In the empirical study, research output is measured by the number of citations of the paper weighted by recursive discounted impact factors of the citing outlet.[18] To capture an author's productivity, we use an author's log lifetime citations (recorded at the beginning of the sample period), decades after receiving their Ph.D., and dummy variables for being a male and a graduate from the Ivy League. Descriptive statistics of the variables of interest can be found in Appendix C.

In Figure 6, we plot the distribution of research output with respect to researcher characteristics. A two-sample t-test indicates that, on average, high-citation researchers have higher output than low-citation researchers, senior researchers have higher output than junior researchers, male researchers have higher output than female researchers, and researchers graduated from Ivy League universities have higher output than researchers graduated from non-Ivy League universities. All results are significant at the 1% level.

4.2 Main Results

In the benchmark empirical model, we assume that the compatibility between researchers is homogeneous, i.e., $g_{ij} = 1$ for $i \neq j$ in Equation (1), and the payoff from a coauthored paper is not discounted, i.e., $\delta_s = 1$ in Equation (2). Table 1 collects the estimation results of Equations (8) and (9), where Column (A) reports the estimates of the production function ignoring endogenous project participation, and Column (B) reports the joint estimates of the production and participation functions with both author- and project-specific random components.

When endogenous project participation is ignored, the estimated complementarity effect

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

$$R_i = \frac{\sum_{j \in \mathcal{J}} R_j C_{ij}}{P_i} \frac{\sum_{j \in \mathcal{J}} P_j}{\sum_{j \in \mathcal{J}} R_j P_j}, \ \forall i \in \mathcal{J}, \tag{10}$$

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

Table 1: Main Results

		(A) Exogenous Participation	(B) Endogenous Participation
Production			
Complementarity	(λ)	-0.0282	0.1164***
		(0.0372)	(0.0151)
Substitutability	(ϕ)	0.0774***	0.2284***
		(0.0289)	(0.0338)
Constant	(β_0)	-1.0076***	-2.8863***
		(0.1714)	(0.1528)
Log life-time citations	(β_1)	0.3069***	0.5577***
		(0.0239)	(0.0209)
Decades after graduation	(β_2)	-0.2128***	-0.4316***
		(0.0437)	(0.0264)
Male	(β_3)	-0.0361	0.0036
		(0.0877)	(0.0398)
Ivy League graduate	(β_4)	0.2538***	0.2981***
		(0.0547)	(0.0325)
Author effect	(ζ)	–	1.6241***
			(0.0584)
Project effect	(ς)	–	0.9708**
			(0.3864)
Error term variance	(σ_ϵ^2)	214.6892***	–
		(5.0884)	
Error term variance	(σ_u^2)	–	91.8476***
			(2.247)
Participation			
Constant	(γ_0)	–	-12.8954***
			(0.2232)
NEP	(γ_1)	–	2.5252***
			(0.1146)
Affiliation	(γ_2)	–	9.0211***
			(0.3359)
Gender	(γ_3)	–	3.5957***
			(0.1717)
Past coauthors	(γ_4)	–	7.6398***
			(0.1566)
Common co-authors	(γ_5)	–	11.6561***
			(0.1879)
Author effect	(ξ)	–	1.6625***
			(0.0857)
Project effect	(ψ)	–	-6.2940***
			(0.1626)
Homophily effect	(κ)	–	-2.0672***
			(0.1076)
Sample size		3,589 papers and 2,421 authors	

Notes: Column (A) estimates the production function ignoring endogenous project participation. Column (B) jointly estimates the production and participation functions with both author and project random effects. We implement MCMC sampling for 30,000 iterations and leave the first 5000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (in the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

(λ) is statistically insignificant; the estimated substitutability effect (ϕ) is statistically significant but its magnitude is small. When endogenous project participation is controlled for, the estimated complementarity effect (λ) becomes significantly positive, and the estimated substitutability effect (ϕ) becomes stronger. A possible explanation for the downward bias of complementarity and substitutability effects is as follows. The estimated coefficients (ζ and ξ) of the author-specific random component suggest that high-ability researchers tend to participate in more projects. Therefore, ignoring endogenous project participation tends to underestimate the substitutability effect because it fails to take into account that the researchers simultaneously working on multiple projects are more likely to be high-ability ones. On the other hand, the estimated coefficients (ς and ψ) of the project-specific random component suggest that high-potential projects (given by η_s) hold a higher threshold for researchers to participate in. The estimated coefficient κ suggests researchers are matched to projects based on homophily of unobserved characteristics. Since high-potential projects are harder to join and researchers are sorted into projects according to their compatibility, high-potential projects with high-ability researchers are relatively scarce in the data. As most researchers in our data are collaborating in projects with relatively low potentials, the complementarity effect is underestimated when endogenous project participation is ignored. In Appendix D, we conduct some Monte Carlo simulation experiments with different signs of ξ and ψ, and the pattern of estimation bias is consistent with the above explanation.

Regarding the effect of author characteristics on research output, we find that the number of lifetime citations is a positive and significant predictor of research output (cf. e.g., Ductor 2015). Being a graduate from an Ivy League university also positively and significantly impacts research output. On the other hand, although Figure 6 shows that senior researchers have higher output on average than junior researchers, the estimation result indicates that, after controlling for network effects and other characteristics, seniority (measured by "decades after graduation") has a negative partial effect on research output.[19] This finding mirrors

[19]Following Rauber & Ursprung (2008) we have also estimated a polynomial of order five in decades after Ph.D. graduation. The result shows that the coefficient of the first order is significantly negative, while those

Ductor (2015), who shows that career time has a negative impact on productivity and it is consistent with the academics' life-cycle effects documented in Levin & Stephan (1991). Similarly, while Figure 6 shows male researchers have higher output on average than female researchers, the estimation result indicates that, after controlling for network effects and other characteristics, gender is not a significant predictor of research output. This result echoes the finding in Ductor et al. (2021) that taking into account coauthorship network significantly reduces the gender output gap.

In the project participation equation, we include the Jaffe similarity[20] between agent i's NEP fields [21] and project s's NEP fields as a measure of the compatibility between the agent and project. We also include covariates that capture the similarities between agent i and the coauthors of project s,[22] to capture the fact that researchers with similar characteristics or past collaborations are more likely to meet and come up with new ideas/projects together. From the estimates reported in Table 1, we find that the similarity in research fields positively and significantly affects the matching between authors and projects (Ductor 2015). In terms of assortative matching between coauthors, belonging to the same affiliation, having the same gender, being coauthors in the past, and sharing common coauthors in the past all make collaboration more likely (cf. Freeman & Huang 2015). In Appendix E, we conduct a robustness check with a participation equation that only includes the similarity between the agent's and project's NEP fields. The main estimation result of the production function is qualitatively unchanged. Therefore, the similarity between the agent's and projects' NEP fields is a leading exogenous factor that controls for the endogenous matching of agents into projects.

of the remaining higher orders are insignificant.

[20]For two vectors, their Jaffe similarity is given by their inner product over the product of their norms. Jaffe (1986) introduces this measure for the analysis of technological similarity between patents. More recently, Bloom et al. (2013) illustrate how "Jaffe similarity" affects firms' profits with different patent portfolios.

[21]We define a researcher's NEP fields based on his/her very first academic publication to alleviate endogeneity concerns.

[22]Take the covariate "affiliation" as an example. Suppose project s has n coauhors. If the number of coauthors of project s (including agent i if he is a coauthor of project s) that belong to the same affiliation as agent i is m, then $\text{affiliation}_{is} = m/n$.

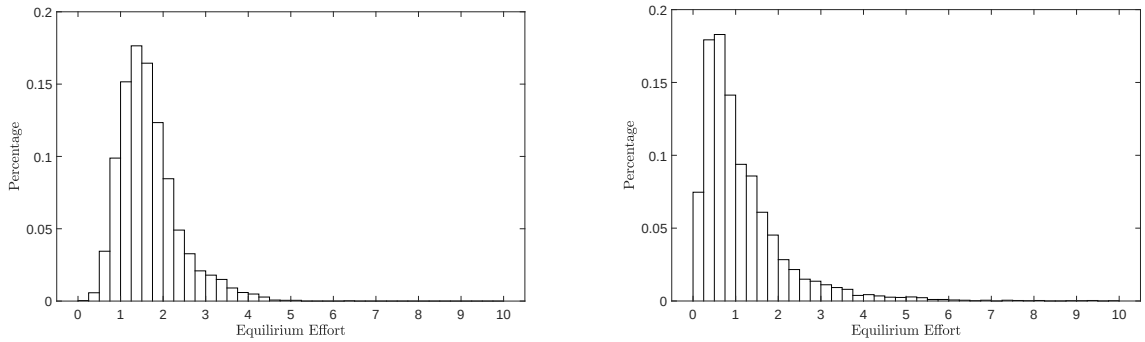

Figure 7: Distributions of equilibrium efforts. The left graph is based on Column (A) of Table 1, and the right graph is based on Column (B) of Table 1.

Finally, in Figure 7, we plot the empirical distributions of equilibrium efforts given by Equation (5) in Proposition 1. We find that the distribution of equilibrium efforts based on Column (B) of Table 1 is more right-skewed than that based on Column (A), suggesting that ignoring endogenous project participation tends to overestimate equilibrium efforts.

4.3 Marginal Effects

From Equation (5), the marginal effect of the kth covariate of agent i on the equilibrium effort is given by

$$\frac{\partial e^*}{\partial x_{ik}} = (I_{np} - L)^{-1} D(\delta \otimes \frac{\partial \alpha}{\partial x_{ik}}),$$

where $\partial \alpha / \partial x_{ik}$ is an $n \times 1$ vector with the ith element being $\partial \alpha_i / \partial x_{ik} = \exp(x_i' \beta)\beta_k$ and other elements being 0.[23] As the agents are connected through the bipartite network, the change in an agent's covariate affects not only his/her own equilibrium effort but also the equilibrium efforts of other agents in the network. The former is known as the *direct* marginal effect, while the latter is known as the *indirect* marginal effect. In Table 2, we report the average marginal effect (AME) of each covariate by first calculating the marginal effect for each individual and then taking an average across all individuals. For the kth covariate, the

[23]The covariate x_{ik} is taken to be a continuous variable. If x_{ik} is a binary variable, then the marginal effect is given by $e^*(x_{ik} = 1) - e^*(x_{ik} = 0)$.

direct AME is given by

$$n^{-1}\sum_{i\in\mathcal{N}}\sum_{s\in\mathcal{P}_i}\frac{\partial e^*_{is}}{\partial x_{ik}},$$

the indirect AME is given by

$$n^{-1}\sum_{i\in\mathcal{N}}\sum_{j\neq i,j\in\mathcal{N},}\sum_{s\in\mathcal{P}_j}\frac{\partial e^*_{js}}{\partial x_{ik}},$$

and the total AME is given by

$$n^{-1}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N},}\sum_{s\in\mathcal{P}_j}\frac{\partial e^*_{js}}{\partial x_{ik}}.$$

The benchmark marginal effects reported in the first column of Table 2 are calculated based on the estimates given in Column (B) of Table 1. To gain a deeper understanding of the magnitudes of the estimated complementarity and substitutability effects, we also calculate the marginal effects under the restrictions $\lambda = 0$, $\phi = 0$, and $\lambda = \phi = 0$ respectively. When the complementarity effect is ignored (i.e., λ is set to 0), the direct AMEs are downward biased by 1%, the indirect AMEs are 0 (i.e., downward biased by 100%), and the total AMEs are downward biased by 8%. When the substitutability effect is ignored (i.e., ϕ is set to 0), the direct AMEs are upward biased by 65%~72%, the indirect AMEs are upward biased by 133%~153%, and the total AMEs are upward biased by 70%~78%. When both effects are ignored (i.e., both λ and ϕ are set to 0), the direct AMEs are upward biased by about 63%~70%, the indirect AMEs are 0, and the total AMEs are upward biased by 51%~58%. In summary, ignoring the complementarity effect leads to a downward bias in the estimated marginal effects, while ignoring the substitutability effect leads to an upward bias in the estimated marginal effects. The latter bias dominates the former.

Table 2: Marginal Effects of Researcher Characteristics on Efforts

	Benchmark	$\lambda = 0$	$\phi = 0$	$\lambda = \phi = 0$
Direct AME				
Log life-time citations	1.4345	1.4208	2.3626	2.3315
Decades after graduation	-1.1072	-1.0966	-1.8227	-1.7988
Male	0.0146	0.0144	0.0251	0.0248
Ivy League graduate	0.8340	0.8259	1.3761	1.3578
Indirect AME				
Log life-time citations	0.1129	0.0000	0.2635	0.0000
Decades after graduation	-0.0871	0.0000	-0.2032	0.0000
Male	0.0011	0.0000	0.0029	0.0000
Ivy League graduate	0.0658	0.0000	0.1545	0.0000
Total AME				
Log life-time citations	1.5475	1.4208	2.6261	2.3315
Decades after graduation	-1.1944	-1.0966	-2.0259	-1.7988
Male	0.0157	0.0144	0.0280	0.0248
Ivy League graduate	0.8998	0.8259	1.5305	1.3578

Notes: The marginal effects are calculated based on the estimates reported in Column (B) of Table 1.

4.4 Robustness Analysis

We also consider two alternative specifications of the empirical model. First, we allow compatibility between researchers to be heterogeneous. Researchers differ in their knowledge bases and these differences can affect their compatibility when collaborating on a joint project. In order to capture heterogeneous compatibility, we define g_{ij} in Equation (1) based on the Jaffe similarity of NEP fields between each pair of authors. The estimation results with heterogeneous compatibility are reported in Table 3. We find the results are comparable with those reported in Table 1. In particular, the complementarity and substitutability effects are both positive and significant when endogenous project participation is controlled for, and are downward biased when endogenous project participation is ignored. It is worth pointing out that the estimates of λ are a little larger than those reported in Table 1. This is because g_{ij} based on the Jaffe similarity measure is smaller than one and thus a larger complementarity coefficient is obtained in compensation.

In the second specification, we assume that the payoff is discounted by the number of coauthors in a project, i.e., $\delta_s = 1/|\mathcal{N}_s|$ in Equation (2).[24] The estimation results are reported in Table 4. Although the estimated complementarity effects are larger than those reported

[24]However, Kuld & O'Hagan (2018) argue that the available empirical evidence suggests that the number of co-authors causes very limited discounting of a published article.

Table 3: Robustness Check: Heterogeneous Compatibility

		(A) Exogenous Participation	(B) Endogenous Participation
Production			
Complementarity	(λ)	-0.0893	0.2059***
		(0.0795)	(0.0217)
Substitutability	(ϕ)	0.0818***	0.2933***
		(0.0284)	(0.0411)
Constant	(β_0)	-0.9902***	-2.8685***
		(0.1670)	(0.1564)
Log life-time citations	(β_1)	0.3092***	0.5812***
		(0.0227)	(0.0219)
Decades after graduation	(β_2)	-0.2135***	-0.5343***
		(0.0430)	(0.0267)
Male	(β_3)	-0.0350	-0.0328
		(0.0867)	(0.0425)
Ivy League graduate	(β_4)	0.2564***	0.3007***
		(0.0535)	(0.0312)
Author effect	(ζ)	–	1.6426***
			(0.0522)
Project effect	(ς)	–	1.4881**
			(0.4379)
Error term variance	(σ_ϵ^2)	214.5831***	–
		(5.0848)	
Error term variance	(σ_u^2)	–	92.0003***
			(2.2335)
Participation			
Constant	(γ_0)	–	-12.4886***
			(0.2095)
NEP	(γ_1)	–	2.4110***
			(0.1089)
Affiliation	(γ_2)	–	8.6645***
			(0.3294)
Gender	(γ_3)	–	3.3041***
			(0.1602)
Past coauthors	(γ_4)	–	7.5885***
			(0.1460)
Common co-authors	(γ_5)	–	11.0446***
			(0.1997)
Author effect	(ξ)	–	1.6079***
			(0.0830)
Project effect	(ψ)	–	-5.3160***
			(0.1607)
Homophily effect	(κ)	–	-1.7855***
			(0.0989)
Sample size		3,589 papers and 2,421 authors	

Notes: Column (A) estimates the production function ignoring endogenous project participation. Column (B) jointly estimates the production and participation functions with both author and project random effects. We implement MCMC sampling for 30,000 iterations and leave the first 5000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (in the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

Table 4: Robustness Check: Discounted Payoffs

		(A) Exogenous Participation	(B) Endogenous Participation
Production			
Complementarity	(λ)	-0.0038	0.3778***
		(0.0961)	(0.0516)
Substitutability	(ϕ)	0.0756***	0.1710***
		(0.0265)	(0.0212)
Constant	(β_0)	-1.0422***	-3.1433***
		(0.1693)	(0.1766)
Log life-time citations	(β_1)	0.3105***	0.5662***
		(0.0231)	(0.0229)
Decades after graduation	(β_2)	-0.2073***	-0.4219***
		(0.0439)	(0.0259)
Male	(β_3)	-0.0352	0.0352
		(0.0870)	(0.0442)
Ivy League graduate	(β_4)	0.2529***	0.3590***
		(0.0535)	(0.0377)
Author effect	(ζ)	–	1.7624***
			(0.0552)
Project effect	(ς)	–	1.2589***
			(0.4416)
Error term variance	(σ_ϵ^2)	214.7029***	–
		(5.0879)	
Error term variance	(σ_u^2)	–	89.4708***
			(2.1938)
Participation			
Constant	(γ_0)	–	-12.1526***
			(0.2201)
NEP	(γ_1)	–	2.3066***
			(0.1132)
Affiliation	(γ_2)	–	8.2854***
			(0.3275)
Gender	(γ_3)	–	3.0565***
			(0.1579)
Past coauthors	(γ_4)	–	7.4929***
			(0.1497)
Common co-authors	(γ_5)	–	10.7551***
			(0.2213)
Author effect	(ξ)	–	1.4193***
			(0.0884)
Project effect	(ψ)	–	-5.4637***
			(0.1659)
Homophily effect	(κ)	–	-1.8165***
			(0.1069)
Sample size		3,589 papers and 2,421 authors	

Notes: Column (A) estimates the production function ignoring endogenous project participation. Column (B) jointly estimates the production and participation functions with both author and project random effects. We implement MCMC sampling for 30,000 iterations and leave the first 5000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (in the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

in Table 1 due to the smaller value of δ_s, the main results are qualitatively unchanged.

In Appendix E, we perform additional robustness checks to gauge the sensitivity of the estimation results. In Table E.1, we experiment with an alternative specification of the participation equation. In Tables E.2 and E.3, we estimate the benchmark empirical model with samples which also include authors who wrote only single-authored papers in the sample period and papers without any citations. We find that the estimates are similar to those reported in Table 1, indicating the robustness of our findings.

4.5 Counterfactual Study

To illustrate the importance of complementarity and substitutability effects in policy evaluation, we conduct a counterfactual study on a simple merit-based research incentives policy. Under this policy, we assume every agent receives merit-based research incentives, $r \in \mathbb{R}_+$, per unit of the output he/she generates.[25] Then the utility function (2) of agent i can be extended to

$$U_i(\mathcal{G}, r) = \sum_{s \in \mathcal{P}_i} (1 + r)\delta_s Y_s - \frac{1}{2}\left(\sum_{s \in \mathcal{P}_i} e_{is}^2 + \phi \sum_{s \in \mathcal{P}_i} \sum_{t \in \mathcal{P}_i \setminus \{s\}} e_{is} e_{it}\right). \tag{11}$$

Let $L(r) := L(r; \lambda, \phi) = \lambda(1 + r)W - \phi M$. Following a similar argument as in the proof of Proposition 1, we can show that, if $\rho_{\max}[L(r)] < 1$, then the equilibrium effort portfolio is given by

$$e^*(r) = (1 + r)[I_{np} - L(r)]^{-1} D(\delta \otimes \alpha). \tag{12}$$

It is worth pointing out that, if the complementarity effect is ignored (i.e., $\lambda = 0$), then $L(r) = -\phi M$, which does not depend on r. In this case, the research incentives r only

[25] Indeed, many universities give awards, monetary incentives, or merit compensation increase to promote high quality research publications.

increase the equilibrium effort

$$e^*(r) = (1+r)(I_{np} + \phi M)^{-1} D(\delta \otimes \alpha)$$

by a factor of $(1+r)$. As the research output is linear in $e^*(r)$ with $\lambda = 0$ in Equation (1), the impact of research incentives on research output is 1:1. Intuitively, when $\lambda = 0$, the multiplier effect of the bipartite network is wiped out, and hence the impact of research incentives is likely to be understated. On the other hand, if the substitutability effect is ignored (i.e., $\phi = 0$), then the cost of research effort is understated and thus the impact of research incentives on equilibrium effort is overstated. As a result, the impact of research incentives on research output tends to be overstated as well.

In Figure 8, the solid line represents the impact of research incentives r on the total research output based on the estimates reported in Column (B) of Table 1.[26] The dashed line corresponds to the case that the complementarity effect is ignored (i.e., λ is set to 0). In this case, the impact of research incentives is understated by about 16%. The dotted line corresponds to the case that the substitutability effect is ignored (i.e., ϕ is set to 0). In this case, the impact of research incentives is overstated by about 90%. When both effects are ignored (i.e., both λ and ϕ are set to 0), the impact of research incentives is depicted by the dash-dotted line. In this case, the impact of research incentives is overstated by about 48%. In summary, consistent with what we observe in the estimated marginal effects, ignoring the complementarity effect underestimates the impact of research incentives, ignoring the substitutability effect overestimates the impact of research incentives, and the latter bias dominates the former. Hence, correctly estimating these two effects is crucial for policy evaluation and recommendation.

[26] δ_s is set to be one for the estimation results reported in Table 1.

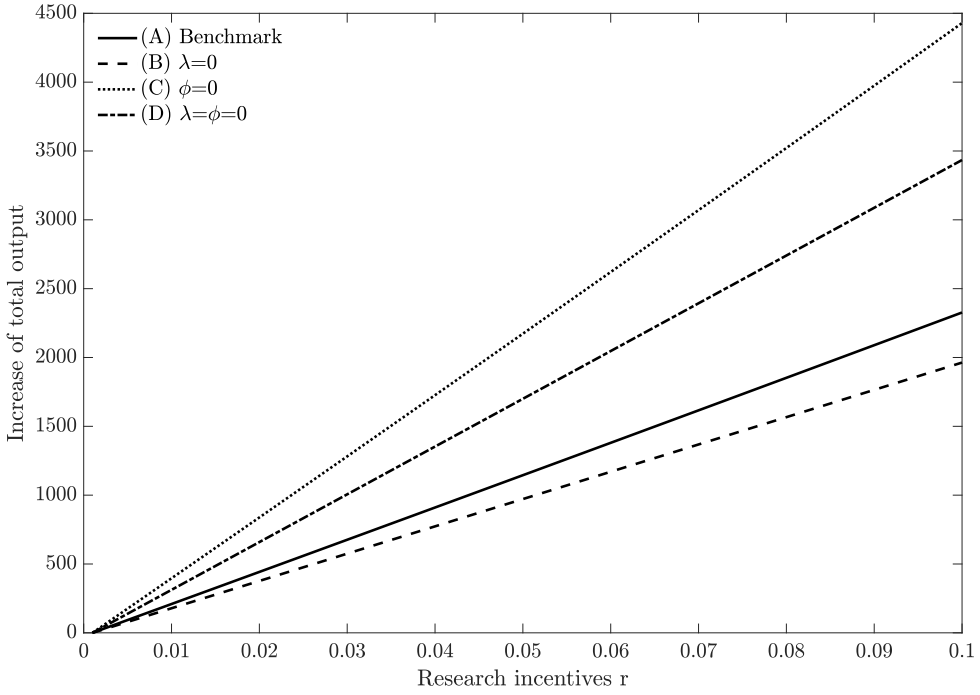

Figure 8: Impact of research incentives on the total research output.

5 Conclusion

In this paper, we analyze the equilibrium efforts of researchers who seek to maximize their utility when involved in multiple, possibly overlapping projects in a bipartite network. We show that both the complementarity effect between collaborating researchers and the substitutability effect between concurrent projects of the same researcher play an important role in determining the equilibrium effort level. To estimate the structural parameters of the model, we propose a Bayesian MCMC procedure that accounts for endogenous selection of researchers into research projects. We then bring our model to the data by analyzing the coauthorship network of economists registered in the RePEc Author Service and find empirical evidence for both complementarity and substitutability effects. Our approach can be applied to other bipartite networks, such as the innovation network of developers and patents and the business network of directors and company boards.

As our model has an explicit micro-foundation, it provides a formal framework for coun-

terfactual analysis. To illustrate the importance of correctly estimating the structural model in policy evaluation, we conduct a counterfactual analysis on the impact of research incentives on research output. We find that the effectiveness of research incentives tends to be underestimated when the complementarity is ignored and overestimated when the substitutability is ignored.

References

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

Anderson, K. A. & Richards-Shubik, S. (2021), Collaborative production in science: An empirical analysis of coauthorships in economics. Forthcoming in *Review of Economics and Statistics*.

Azoulay, P., Graff Zivin, J. S. & Wang, J. (2010), 'Superstar extinction', *The Quarterly Journal of Economics* **125**(2), 549–589.

Ballester, C., Calvó-Armengol, A. & 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. & Deroïan, F. (2014), 'Competing activities in social networks', *The BE Journal of Economic Analysis & Policy* **14**(4), 1431–1466.

Bimpikis, K., Ehsani, S. & Ilkilic, R. (2019), 'Cournot competition in networked markets', *Management Science* **65**(6), 2467–2481.

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

Bonhomme, S. (2020), Econometric analysis of bipartite networks, *in* B. Graham &

A. de Paula, eds, 'The Econometric Analysis of Network Data', Academic Press, pp. 83–121.

Bonhomme, S. (2021), Teams: Heterogeneity, sorting, and complementarity. Working paper, University of Chicago.

Bramoullé, Y., Djebbari, H. & Fortin, B. (2009), 'Identification of peer effects through social networks', *Journal of Econometrics* **150**(1), 41–55.

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

Colussi, T. (2017), 'Social ties in academia: A friend is a treasure', *Review of Economics and Statistics* **100**(1), 45–50.

Currarini, S., Jackson, M. & Pin, P. (2009), 'An economic model of friendship: Homophily, minorities and segregation', *Econometrica* **77**(4), 1003–1045.

Ductor, L. (2015), 'Does co-authorship lead to higher academic productivity?', *Oxford Bulletin of Economics and Statistics* **77**(3), 385–407.

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

Ductor, L., Goyal, S. & Prummer, A. (2021), Gender and collaboration. Forthcoming in *Review of Economics and Statistics*.

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

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

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

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

Jackson, M. O. & 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), 984–1001.

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

Koop, G., Poirier, D. J. & Tobias, J. L. (2007), *Bayesian Econometric Methods*, Cambridge University Press.

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

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

Lindenlaub, I. & Prummer, A. (2021), 'Network structure and performance', *The Economic Journal* **131**(634), 851–898.

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

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001), 'Birds of a feather: Homophily in social networks', *Annual Review of Sociology* **27**, 415–444.

Newman, M. E. J. (2001*a*), 'Scientific collaboration networks. I. Network construction and fundamental results', *Physical Review E* **64**(1), 016131.

Newman, M. E. J. (2001*b*), 'Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality', *Physical Review E* **64**(1), 016132.

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

Newman, M. E. J. (2004*a*), 'Coauthorship networks and patterns of scientific collaboration', *Proceedings of the National Academy of Sciences* **101**(suppl 1), 5200–5205.

Newman, M. E. J. (2004*b*), Who is the best connected scientist? a study of scientific coauthorship networks, *in* E. Ben-Naim, H. Frauenfelder & Z. Toroczkai, eds, 'Complex Net-

works. Lecture Notes in Physics', Vol. 650, Springer, Berlin, Heidelberg, pp. 337–370.

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

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

Samelson, H., Thrall, R. M. & Wesler, O. (1958), 'A partition theorem for Euclidean n-space', *Proceedings of the American Mathematical Society* **9**(5), 805–807.

Singh, N. & Vives, X. (1984), 'Price and quantity competition in a differentiated duopoly', *The RAND Journal of Economics* **15**(4), 546–554.

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

Zeger, S. L. & Karim, M. R. (1991), 'Generalized linear models with random effects; A Gibbs sampling approach', *Journal of the American Statistical Association* **86**(413), 79–86.

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

Appendices for "Collaboration in Bipartite Networks"

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

A Proof of Proposition 1

Proof of Proposition 1. Let d_{is} be an indicator variable such that $d_{is} = 1$ if agent i participates in project s and $d_{is} = 0$ otherwise. Substitution of Equation (1) into Equation (2) gives

$$\begin{aligned} U_i(\mathcal{G}) &= \sum_{s\in\mathcal{P}} d_{is}\delta_s \left(\sum_{j\in\mathcal{N}} \alpha_j d_{js} e_{js} + \frac{\lambda}{2} \sum_{j\in\mathcal{N}} \sum_{k\in\mathcal{N}\setminus\{j\}} g_{jk} d_{js} d_{ks} e_{js} e_{ks} + \epsilon_s \right) \\ &\quad -\frac{1}{2} \left(\sum_{s\in\mathcal{P}} d_{is} e_{is}^2 + \phi \sum_{s\in\mathcal{P}} \sum_{t\in\mathcal{P}\setminus\{s\}} d_{is} d_{it} e_{is} e_{it} \right). \end{aligned} \tag{13}$$

First, note that the marginal utility has to be non-positive at equilibrium, i.e.,

$$\frac{\partial U_i(\mathcal{G})}{\partial e_{is}}|_{e^*} = d_{is} \left(\delta_s \alpha_i + \lambda \delta_s \sum_{j\in\mathcal{N}\setminus\{i\}} g_{ij} d_{js} e_{js}^* - e_{is}^* - \phi \sum_{t\in\mathcal{P}\setminus\{s\}} d_{it} e_{it}^* \right) \leq 0,$$

where the inequality is strict only if $e_{is}^* = 0$ at equilibrium (corner solution). This set of inequalities can be written in matrix form as

$$-D(\delta \otimes \alpha) + (I_{np} - L)e^* \geq 0. \tag{14}$$

Second, if $e_{is}^* > 0$ at equilibrium, then $\frac{\partial U_i(\mathcal{G})}{\partial e_{is}}|_{e^*} = 0$, which implies

$$e^{*\prime}[-D(\delta \otimes \alpha) + (I_{np} - L)e^*] = 0. \tag{15}$$

Finally, the equilibrium effort has to be non-negative, i.e.,

$$e^* \geq 0. \tag{16}$$

Conditions (14), (15), and (16) constitute a linear complementarity problem (Samelson et al. 1958). If $\rho_{\max}(L) < 1$, the matrix $I_{np} - L$ is positive definite. It follows by Lemmas 2 and 3 in Bimpikis et al. (2019) that the unique equilibrium is given by the solution to the linear complementarity problem and the inactive links ($d_{is} = 0$) are strategically redundant and play no role in determining the equilibrium. Hence, it follows by a similar arguments as in the proof of Theorem 1 in Bimpikis et al. (2019) that the game has a unique equilibrium with the equilibrium effort levels are given by Equation (5). ∎

B Bayesian Estimation

Since the likelihood function based on Equations (8) and (9) involves high-dimensional integrals, it is computationally cumbersome to apply a frequentist maximum likelihood method even when resorting to a simulation approach. As an alternative estimation method, the Bayesian Markov Chain Monte Carlo (MCMC) approach can be more efficient for estimating latent variable models (cf. Zeger & Karim 1991). We divide the parameter vector θ and unknown latent variables into blocks and assign the prior distributions as follows:

$$\begin{aligned}
\lambda &\sim \mathcal{N}(0, \sigma_\lambda^2), \\
\phi &\sim \mathcal{N}(0, \sigma_\phi^2), \\
\beta &\sim \mathcal{N}(0, \Sigma_\beta), \\
\zeta &\sim \mathcal{N}(0, \sigma_\zeta^2), \\
\varsigma &\sim \mathcal{N}(0, \sigma_\varsigma^2), \\
\gamma &\sim \mathcal{N}(0, \Sigma_\gamma), \\
\xi &\sim \mathcal{N}(0, \sigma_\xi^2), \\
\psi &\sim \mathcal{N}(0, \sigma_\psi^2), \\
\kappa &\sim \mathcal{N}(0, \sigma_\kappa^2), \\
\sigma_u^2 &\sim \mathcal{IG}\left(\frac{\tau_0}{2}, \frac{\nu_0}{2}\right),
\end{aligned}$$

and $\mu_i \sim \mathcal{N}(0,1)$ for $i \in \mathcal{N}$ and $\eta_s \sim \mathcal{N}(0,1)$ for $s \in \mathcal{P}$. We consider the normal and inverse gamma ($\mathcal{IG}$) conjugate priors, which are widely used in the Bayesian literature (Koop et al. 2007). The hyperparameters are chosen to make the prior distributions relatively flat and cover a wide range of the parameter space, i.e., we set $\sigma_\lambda^2 = \sigma_\phi^2 = 10$, $\Sigma_\beta = 10I$, $\sigma_\zeta^2 = \sigma_\varsigma^2 = 10$, $\Sigma_\gamma = 1000I$, $\sigma_\xi^2 = \sigma_\psi^2 = \sigma_\kappa^2 = 1000$, $\tau_0 = 2.2$, and $\nu_0 = 0.1$.

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

1. Draw the latent variable μ_i using the M-H algorithm based on $f(\mu_i|y, d, \theta, \mu_{-i}, \eta)$, for $i = 1, \ldots, n$.

2. Draw the latent variable η_s using the M-H algorithm based on $f(\eta_s|y, d, \theta, \mu, \eta_{-s})$, for $s = 1, \ldots, p$.

3. Draw γ using the M-H algorithm based on $f(\gamma|y, d, \theta\backslash\{\gamma\}, \mu, \eta)$.

4. Draw ξ using the M-H algorithm based on $f(\xi|y, d, \theta\backslash\{\xi\}, \mu, \eta)$.

5. Draw ψ using the M-H algorithm based on $f(\psi|y, d, \theta\backslash\{\psi\}, \mu, \eta)$.

6. Draw κ using the M-H algorithm based on $f(\kappa|y, d, \theta\backslash\{\kappa\}, \mu, \eta)$.

7. Draw λ using the M-H algorithm based on $f(\lambda|y, d, \theta\backslash\{\lambda\}, \mu, \eta)$.

8. Draw ϕ using the M-H algorithm based on $f(\phi|y, d, \theta\backslash\{\phi\}, \mu, \eta)$.

9. Draw β using the M-H algorithm based on $f(\beta|y, d, \theta\backslash\{\beta\}, \mu, \eta)$.

10. Draw ζ using the M-H algorithm based on $f(\zeta|y, d, \theta\backslash\{\zeta\}, \mu, \eta)$.

11. Draw ς using the M-H algorithm based on $f(\varsigma|y, d, \theta\backslash\{\varsigma\}, \mu, \eta)$.

12. Draw σ_u^2 using the conjugate inverse gamma conditional posterior distribution.

We collect the draws from iterating the above steps and compute the posterior mean and the posterior standard deviation as our estimation results.

C Data Description

To obtain a sample from RePEc that is appropriate for our analysis, we apply a series of filters as follows.

First, we select papers that had a first pre-print version in 2010-2012. Furthermore, we require all authors of the papers to be registered with RePEc and all authors to have the RePEc Genealogy information on where they studied. We drop all duplicate or older versions of each paper from our sample. This gives us a sample of 6,673 papers written by 3,700 distinct authors for which we have complete data. We call this sample: Sample (I). This is the sample that we used to obtain the estimates reported in Table E.2 in Appendix E. Descriptive statistics of the variables of interest in Sample (I) are reported in Table C.1.

Next, we drop 2,463 papers that do not have any citations up to July 2018 when the data is extracted from the RePEc database, as well as 658 authors who only work on these dropped papers without any citations. This reduces the sample size to 4,210 papers and 3,042 authors. We call this sample: Sample (II). This is the sample that we used to obtain the estimates reported in Table E.3 in Appendix E. Descriptive statistics of the variables of interest in Sample (II) are reported in Table C.2.

Finally, we drop 621 authors who only wrote a single-authored paper in the sample period. This results in a sample of 3,589 papers written by 2,421 distinct authors. We call this sample: Sample (III). This is the sample we used to obtain the main results reported in Section 4.2. Descriptive statistics of the variables of interest in Sample (III) are reported in Table C.3.

Table C.1: Summary Statistics of Sample (I)

	Min	Max	Mean	S.D.	Sample size
Papers					
Weighted citations	0.0000	317.9515	3.7587	12.3858	6673
Number of authors (in each paper)	1	5	1.4160	0.6421	6673
Authors					
Log life-time citations	0	10.7634	5.3176	1.8428	3700
Decades after graduation	-0.7	6.2000	1.0642	1.0676	3700
Male	0	1	0.8154	0.3880	3700
Ivy League graduate	0	1	0.1268	0.3327	3700
Number of papers (for each author)	1	63	2.5538	2.7762	3700

Notes: This sample is constructed based on works that were released as working papers in 2010-2012. We drop papers in which not all of their authors were registered with RePEc. We also drop authors who do not have the RePEc Genealogy information on where they studied.

Table C.2: Summary Statistics of Sample (II)

	Min	Max	Mean	S.D.	Sample size
Papers					
Weighted citations	0.0000	317.9515	5.9577	15.1682	4210
Number of authors (in each paper)	1	5	1.5124	0.6820	4210
Authors					
Log life-time citations	0	10.7634	5.5445	1.7358	3042
Decades after graduation	-0.7	6.2000	1.0701	1.0447	3042
Male	0	1	0.8222	0.3824	3042
Ivy League graduate	0	1	0.1341	0.3408	3042
Number of papers (for each author)	1	19	2.0930	1.7079	3042

Notes: This sample is constructed based on works that were released as working papers in 2010-2012. We drop papers in which not all of their authors were registered with RePEc. We also drop authors who do not have the RePEc Genealogy information on where they studied. In this sample, we further drop papers which do not have any citations up to November 2018.

Table C.3: Summary Statistics of Sample (III)

	Min	Max	Mean	S.D.	Sample size
Papers					
Weighted citations	1e-04	317.9515	6.4578	16.0868	3589
Number of authors (in each paper)	1	5	1.6010	0.7017	3589
Authors					
Log life-time citations	0	10.7634	5.7441	1.6782	2421
Decades after graduation	-0.7	6.2000	1.1056	1.0372	2421
Male	0	1	0.8228	0.3819	2421
Ivy League graduate	0	1	0.1450	0.3522	2421
Number of papers (for each author)	1	19	2.3734	1.8112	2421

Notes: This sample is constructed based on works that were released as working papers in 2010-2012. We drop papers in which not all of their authors were registered with RePEc. We also drop authors who do not have the RePEc Genealogy information on where they studied. In this sample, we further drop papers which do not have any citations up to November 2018 and the authors who only wrote a single-authored paper in the sample period.

D Monte Carlo Simulation

To show that the proposed Bayesian MCMC estimation approach in Appendix B can effectively recover the true parameters in Equations (8) and (9), we conduct a Monte Carlo simulation with 100 repetitions. In each repetition, we generate an artificial bipartite collaboration network of 300 authors ($n = 300$) and 400 projects ($p = 400$). The data generating process (DGP) runs as follows: we first simulate dyadic binary exogenous variables $z_{is} \in \{0, 1\}$ randomly with the probability $P(z_{is} = 1) = 0.64$; individual exogenous variable x_i from normal distribution $N(0, 4)$; and both author and project latent variables μ_i and η_s from $N(0, 1)$. Then, we generate the artificial collaboration network and project output based on the participation function of Equation (9) and the production function of Equation (8).

In the Monte Carlo simulations, we consider three sets of parameters to see how the signs of the coefficients of agent and project latent variables affect the direction of the selection bias. In the first parameter specification, we set $\zeta > 0$ and $\xi > 0$ (i.e., a researcher with higher ability μ_i tends to participate in more projects), $\varsigma > 0$ and $\psi < 0$ (i.e., a project with higher potential η_s has a higher threshold for researchers to participate in), and $\kappa < 0$

(i.e., agents are sorted into projects based on homophily of unobserved characteristics). The simulation results reported in Table D.1 confirm that all true model parameters can be effectively recovered by the employed Bayesian MCMC approach when endogenous project participation is controlled for through author and project latent variables, and both the complementarity and substitutability effects are downward biased when endogenous project participation is ignored. The direction of the bias is the same as what we observe in the empirical study.

In the second parameter specification, we set $\zeta > 0$ and $\xi < 0$ (i.e., a researcher with higher ability tends to participate in fewer projects), while holding the other parameters the same as the first specification. In this case, all true model parameters can still be effectively recovered by the employed Bayesian MCMC approach when endogenous project participation is controlled for. When endogenous project participation is ignored, the substitutability effect is overestimated because it is low-ability researchers who are more likely to work on multiple concurrent projects.

In the third parameter specification, we set $\varsigma > 0$ and $\psi > 0$ (i.e., high-potential projects are easier to join than low-potential ones), while holding the other parameters the same as the first specification. In this case, all true model parameters can still be effectively recovered by the employed Bayesian MCMC approach when endogenous project participation is controlled for. When endogenous project participation is ignored, the complementarity effect is overestimated because researchers are more likely to collaborate on high-potential projects.

From the simulation results, we can conclude (i) all true model parameters can be effectively recovered by the employed Bayesian MCMC approach when endogenous project participation is controlled for, and (ii) the pattern of bias is consistent with our intuition.

Table D.1: Simulation results: downward biases on λ and ϕ

			Exogenous Participation		Endogenous Participation	
		DGP	Est.	S.D.	Est.	S.D.
Production						
Complementarity	(λ)	0.10	0.0315	0.0609	0.0997	0.0024
Substitutability	(ϕ)	0.10	-0.1051	0.1171	0.1007	0.0146
Constant	(β_0)	-1.00	-0.0584	0.6452	-0.9965	0.0620
x_i	(β_1)	0.50	0.3877	0.1754	0.4974	0.0131
Author effect	(ζ)	1.00			1.0014	0.0262
Project effect	(ς)	0.50			0.5251	0.0459
Error variance	(σ_u^2)	0.50	236.3430	176.4898	0.4625	0.0438
Participation						
Constant	(γ_0)	-5.75			-5.7298	0.1008
z_{ij}	(γ_1)	0.50			0.4882	0.1041
Author effect	(ξ)	1.00			1.2646	0.2290
Project effect	(ψ)	-1.00			-1.2657	0.2538
Homophily	(κ)	-0.50			-0.7721	0.2373

Notes: We perform Monte Carlo simulations with 100 repetitions. The reported values are the mean and the standard deviation of point estimates calculated across repetitions.

Table D.2: Simulation Result: upward bias on ϕ

			Exogenous Participation		Endogenous Participation	
		DGP	Est.	S.D.	Est.	S.D.
Production						
Complementarity	(λ)	0.10	0.0539	0.0509	0.0934	0.0271
Substitutability	(ϕ)	0.10	0.8949	1.1015	0.0997	0.0342
Constant	(β_0)	-1.00	-0.6415	0.9804	-0.9829	0.0967
x_i	(β_1)	0.50	0.4004	0.3622	0.4913	0.0250
Author effect	(ζ)	1.00			1.0042	0.0507
Project effect	(ς)	0.50			0.5259	0.0360
Error variance	(σ_u^2)	0.50	77.3531	76.722	0.4753	0.0354
Participation						
Constant	(γ_0)	-5.75			-5.7416	0.1063
z_{ij}	(γ_1)	0.50			0.4964	0.0971
Author effect	(ξ)	-1.00			-1.2362	0.2490
Project effect	(ψ)	-1.00			-1.2485	0.2418
Homophily	(κ)	-0.50			-0.7283	0.2754

Notes: We perform Monte Carlo simulations with 100 repetitions. The reported values are the mean and the standard deviation of point estimates calculated across repetitions.

Table D.3: Simulation Result: upward bias on λ

			Exogenous Participation		Endogenous Participation	
		DGP	Est.	S.D.	Est.	S.D.
Production						
Complementarity	(λ)	0.10	0.1277	0.0096	0.1005	0.0015
Substitutability	(ϕ)	0.10	-0.0197	0.0720	0.0968	0.0039
Constant	(β_0)	-1.00	-1.4053	1.2002	-1.0625	0.0338
x_i	(β_1)	0.50	0.5845	0.3356	0.5077	0.0124
Author effect	(ζ)	1.00			0.9957	0.0221
Project effect	(ς)	0.50			0.4992	0.0460
Error variance	(σ_u^2)	0.50	904.6899	752.5478	0.5111	0.0412
Participation						
Constant	(γ_0)	-5.75			-5.7572	0.0894
z_{ij}	(γ_1)	0.50			0.4839	0.0938
Author effect	(ξ)	1.00			0.9867	0.0488
Project effect	(ψ)	1.00			0.9734	0.0540
Homophily	(κ)	-0.50			-0.5036	0.0656

Notes: We perform Monte Carlo simulations with 100 repetitions. The reported values are the mean and the standard deviation of point estimates calculated across repetitions.

E Additional Robustness Checks

In this section, we perform additional robustness checks to gauge the sensitivity of the estimation results. In Table E.1, we experiment with an alternative specification of the participation equation. In Tables E.2 and E.3, we estimate the benchmark empirical model with Sample (I) and Sample (II) respectively (see Appendix C). We find that the estimates are similar to those reported in Table 1, indicating the robustness of our findings.

Table E.1: Robustness Check: Alternative Participation Equations

		(A) Homogeneous Complementarity	(B) Heterogeneous Complementarity	(C) Discounted Payoffs
Production				
Complementarity	(λ)	0.1110***	0.2259***	0.3556***
		(0.0159)	(0.0196)	(0.0448)
Substitutability	(ϕ)	0.2041***	0.2311***	0.1781***
		(0.0285)	(0.0311)	(0.0212)
Constant	(β_0)	-2.9599***	-2.5822***	-3.0756***
		(0.1535)	(0.1423)	(0.1490)
Log life-time citations	(β_1)	0.5580***	0.4978***	0.5599***
		(0.0208)	(0.0212)	(0.0199)
Decades after graduation	(β_2)	-0.4315***	-0.3847***	-0.4307***
		(0.0255)	(0.0275)	(0.0248)
Male	(β_3)	-0.0111	-0.0279	0.0464
		(0.0403)	(0.0457)	(0.0387)
Ivy League graduate	(β_5)	0.3222***	0.3665***	0.3344***
		(0.0303)	(0.0458)	(0.0278)
Author effect	(ζ)	1.7854***	1.6729***	1.7871***
		(0.0550)	(0.0520)	(0.0528)
Project effect	(ς)	1.0686**	1.1877**	1.4627***
		(0.4566)	(0.4246)	(0.4302)
Error term variance	(σ_u^2)	88.8616***	86.8997***	88.9253***
		(2.4089)	(2.1576)	(2.1357)
Participation				
Constant	(γ_0)	-7.7101***	-7.6576***	-7.7143***
		(0.0542)	(0.0845)	(0.0875)
NEP	(γ_1)	1.6197***	1.6131***	1.6129***
		(0.0945)	(0.0997)	(0.0978)
Author effect	(ξ)	0.4461***	0.5163***	0.4406***
		(0.0628)	(0.0632)	(0.0635)
Project effect	(ψ)	-0.5582***	-0.6755***	-0.5382***
		(0.0944)	(0.0887)	(0.0817)
Homophily effect	(κ)	-0.7438***	-0.7466***	-0.7471***
		(0.0738)	(0.0690)	(0.0685)
Sample size		3,589 papers and 2,421 authors		

Notes: Column (A) assumes homogeneous complementarity. Column (B) allows for heterogeneous complementarity using Jaffe's similarity measure for the research fields of collaborating authors. Column (C) considers the case where the payoff is discounted by the number of coauthors in a project. We implement MCMC sampling for 30,000 iterations and leave the first 5000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (in the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

Table E.2: Robustness Check: Sample (I)

		(A) Exogenous Participation	(B) Endogenous Participation
Production			
Complementarity	(λ)	0.0021	0.1598***
		(0.0260)	(0.0116)
Substitutability	(ϕ)	0.1014***	0.1441***
		(0.0233)	(0.0227)
Constant	(β_0)	-1.2066***	-3.4528***
		(0.1256)	(0.1437)
Log life-time citations	(β_1)	0.3303***	0.6186***
		(0.0175)	(0.0221)
Decades after graduation	(β_2)	-0.2371***	-0.5033***
		(0.0335)	(0.0257)
Male	(β_3)	-0.0288	-0.1442**
		(0.0700)	(0.0557)
Ivy League graduate	(β_5)	0.2919***	0.3851***
		(0.0458)	(0.0396)
Author effect	(ζ)	–	2.0052***
			(0.0655)
Project effect	(ς)	–	-0.6477
			(0.4203)
Error term variance	(σ_ϵ^2)	126.7968***	–
		(2.1856)	
Error term variance	(σ_u^2)	–	72.2692***
			(1.4150)
Participation			
Constant	(γ_0)	–	-10.6088***
			(0.1233)
NEP	(γ_1)	–	1.4979***
			(0.0771)
Affiliation	(γ_2)	–	6.5643***
			(0.2446)
Gender	(γ_3)	–	1.7358***
			(0.0841)
Past coauthors	(γ_4)	–	6.2987***
			(0.0948)
Common co-authors	(γ_5)	–	7.2021***
			(0.0729)
Author effect	(ξ)	–	0.4572***
			(0.0582)
Project effect	(ψ)	–	-1.2853***
			(0.0919)
Homophily effect	(κ)	–	-0.2246***
			(0.0831)
Sample size		6,673 papers and 3,700 authors	

Notes: Column (A) estimates the production function ignoring endogenous project participation. Column (B) jointly estimates the production and participation functions with both author and project random effects. We implement MCMC sampling for 30,000 iterations and leave the first 5000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (in the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.

Table E.3: Robustness Check: Sample (II)

		(A) Exogenous Participation	(B) Endogenous Participation
Production			
Complementarity	(λ)	-0.0407	0.1422***
		(0.0357)	(0.0118)
Substitutability	(ϕ)	0.0827***	0.1847***
		(0.0254)	(0.0123)
Constant	(β_0)	-0.9694***	-3.1811***
		(0.1480)	(0.1454)
Log life-time citations	(β_1)	0.3121***	0.6027***
		(0.0215)	(0.0200)
Decades after graduation	(β_2)	-0.2334***	-0.5232***
		(0.0416)	(0.0216)
Male	(β_3)	-0.0501	-0.0008
		(0.0766)	(0.0339)
Ivy League graduate	(β_5)	0.2613***	0.2994***
		(0.0499)	(0.0247)
Author effect	(ζ)	–	1.6756***
			(0.0481)
Project effect	(ς)	–	1.5924***
			(0.2942)
Error term variance	(σ_ϵ^2)	189.6873***	–
		(4.1484)	
Error term variance	(σ_u^2)	–	76.9183***
			(1.6947)
Participation			
Constant	(γ_0)	–	-12.5956***
			(0.2253)
NEP	(γ_1)	–	2.3455***
			(0.1095)
Affiliation	(γ_2)	–	8.9033***
			(0.3213)
Gender	(γ_3)	–	3.3093***
			(0.1363)
Past coauthors	(γ_4)	–	7.4911***
			(0.1635)
Common co-authors	(γ_5)	–	10.9856***
			(0.1859)
Author effect	(ξ)	–	1.5721***
			(0.0730)
Project effect	(ψ)	–	-4.2284***
			(0.1082)
Homophily effect	(κ)	–	-1.6136***
			(0.0792)
Sample size		4,210 papers and 3,042 authors	

Notes: Column (A) estimates the production function ignoring endogenous project participation. Column (B) jointly estimates the production and participation functions with both author and project random effects. We implement MCMC sampling for 30,000 iterations and leave the first 5000 draws for burn-in and use the rest of draws for computing the posterior mean (as the point estimate) and the posterior standard deviation (in the parenthesis). The asterisks ***(**,*) indicate that the 99% (95%, 90%) highest posterior density range does not cover zero.