Collaborative Production in Science:

An Empirical Analysis of Coauthorships in Economics

Katharine A. Anderson (Interos)

Seth Richards-Shubik (Lehigh University and NBER)

March 3, 2021

Abstract

This paper studies productivity and preferences in scientific research. Collaboration is increasingly important for innovation in science, and other domains, but we have limited understanding of the factors researchers use to choose their collaborators and the projects they work on. Here, we use a model of strategic network formation and a recently developed econometric method to examine this question in the context of economics researchers. We learn that research teams with more collaborators tend to produce papers with higher impact, without increasing individual costs of communication and coordination. This suggests the trend toward larger research teams in economics will continue.

Keywords: collaborative production, scientific research, coauthorship network, economics publishing, network formation, structural estimation, partial identification

JEL Codes: L14, D85, O31, A14, C57, C55

1 Introduction

Collaborative production is an area of increasing economic importance. It allows workers to combine their skills in new ways, both increasing productivity and speeding the pace of innovation (e.g., Hong and Page, 2001; Polzer, Milton and Swann, 2002; Hamilton, Nickerson and Owan, 2003). In particular, scientific production has grown increasingly collaborative, and so the analysis of scientific productivity and innovation must take this into account (Moody, 2004; Goyal, van der Leij and Moraga-Gonzalez, 2006; Wuchty, Jones and Uzzi, 2007; Hamermesh, 2013; Card and DellaVigna, 2013; Anderson, Crespi and Sayre, 2017).

The relationships among researchers can be captured using a collaboration network: a web of interactions in which two participants are connected if they work together on a project. The literature suggests that the structure of these networks has a real effect on the both the outcomes of individuals, and the overall pace of innovation. Deeply interlinked parts of the network allow for more effective communication, while connections between those interlinked areas allow for cross-pollination (Burt, 2001; Moody, 2004). Moreover, the outcome for any particular individual potentially depends on their connections and the network structure around them (Katz and Hicks, 1997; Ductor et al., 2014).

In order to maximize their chances of success, researchers spend considerable time deciding which projects to pursue and which colleagues to pursue them with. These decisions are based both on the potential impact of the research and the cost of pursuing it. Given that the individual connections are the result of a very careful decision-making process, we can derive information about those decisions from the observed collaboration network.

We use a model of academic collaborative decisions, in combination with the observed collaboration network in economics, to learn about the underlying motivations of researchers in this discipline. In our model, researchers decide on a combination of paper topic and set of collaborators, based on the expected quality of the publication—in the form of a journal

impact score—and the costs of producing that paper with that group of collaborators. The authors jointly make their decisions, and each collaboration is represented by a set of links in the collaboration network.

The literature has examined many factors that may affect the quality and the cost of research production, which researchers might consider when making their decisions. These include the number of authors, the match between their expertise and the chosen topic, and their overall ability, perhaps reflected in their prior work (e.g., Wuchty, Jones and Uzzi, 2007; Ductor, 2015; Azoulay, Graff Zivin and Wang, 2010). The quality and impact of a research project may also be a function of the authors' connections in the broader collaboration network (e.g., Uddin, Hossain and Rasmussen, 2013; Ductor et al., 2014; Colussi, 2018). The literature has also highlighted potential communication and coordination difficulties within teams, that relate to the size of a group and the specialization of its members (e.g., Becker and Murphy, 1992; Dessein and Santos, 2006; Cremer, Garicano and Prat, 2007; Walsh and Maloney, 2007). Studies on interdisciplinary research additionally indicate there may be costs from working on more novel and unfamiliar topics (e.g., Rhoten and Parker, 2004; Leahey, Beckman and Stanko, 2017).

Our empirical approach, described next, enables us to evaluate some but not all of these factors. In particular, we do not address the role of "star" researchers, as in Goyal, van der Leij and Moraga-Gonzalez (2006) and Azoulay, Graff Zivin and Wang (2010). Instead, our analysis focuses on the productivity and preferences of "regular" economists, who comprise the vast majority of authors. Among the factors we assess, our strongest results pertain to the number of authors on research papers, a topic of ongoing interest in economics (McDowell and Melvin, 1983; Barnett, Ault and Kaserman, 1988; Hamermesh, 2013; Card and DellaVigna, 2013; Ellison, 2013; Rath and Wohlrabe, 2016; Kuld and O'Hagan, 2018). Given that we observe collaborations of increasing size, a natural question is what motivates this, and what

are the trade-offs that researchers face when choosing larger teams. Our results indicate that research teams with more collaborators tend to produce papers with higher impact, without increasing individual costs of communication and coordination.

To recover the benefits and costs that researchers consider when choosing their collaborations, we structurally estimate an equilibrium model of network formation. This approach is computationally intensive, but it is necessary given the joint nature of the choices that produce the observed collaborations. We use a recently developed framework by de Paula, Richards-Shubik and Tamer (2018, abbreviated "PRT"), which yields a set of structural parameter vectors that could be consistent with the observed network. This framework requires a strict bound on the degree of each node, making it better suited for regular researchers with relatively few papers, rather than star researchers. Also, because the method is computationally intensive—as are other methods for strategic network formation models—we must limit the number of factors in the model and use highly parsimonious specifications.

To adapt the econometric framework to our application and make it tractable, we introduce several substantive extensions. First, our network is a multigraph, because researchers may have multiple projects together, and the edges have attributes that represent each paper's topic. We accordingly extend the PRT framework that was developed for simple graphs. Second, our model has a production function (for paper impact) embedded within the utility function, which can be estimated separately, prior to the utility parameters. This enables us to include a relatively large number of production factors without greatly increasing the computational burden. Third, we collect individuals with very rare outcomes into residual categories, in order to reduce the computational burden. Last, we develop a computationally efficient method to incorporate sampling uncertainty in the observed outcomes.

The model is estimated with data from EconLit, a comprehensive database of articles in economics journals. We use publications in 2009 and 2010 to define a single, static,

outcome period. Prior publications from 1998 to 2007 are also used to provide information on researcher skills and productivity. We restrict the analysis to "experienced" authors, meaning those who published at least two articles in the earlier period. Papers with both experienced and inexperienced authors are included, but we examine the preferences of experienced researchers only. Our results, as noted earlier, indicate that adding coauthors both increases quality and reduces individual costs, on average. This is not obvious because, as argued by Becker and Murphy (1992) and Dessein and Santos (2006), among others, coordination costs can limit optimal team size and specialization. In addition, we do not find strong evidence that differences in skill backgrounds increase communication costs, or that generalists mitigate any such costs. Taken together, these results suggest that the trend in economics toward larger research teams with more diverse skill backgrounds will continue.

Our study adds to a large and growing literature on collaboration and productivity in economic research. This literature has provided sophisticated descriptive analyses of the collaboration network (Goyal, van der Leij and Moraga-Gonzalez, 2006) and the observed relationships between team size, composition, and external connections on publication outcomes (Bosquet and Combes, 2013; Ductor et al., 2014; Ductor, 2015; Colussi, 2018).[1] By structurally estimating an equilibrium model, we reveal the (possibly) fixed incentive structures that influence the choices behind these observed patterns. A closely related study by Hsieh et al. (2018) structurally estimates a model of effort allocation among coauthors, given exogenously established collaborations, and similarly uses data on economics papers.[2]

[1]There are also important results on homophily by gender and ethnicity (e.g., Boschini and Sjögren, 2007; Freeman and Huang, 2015), but our model does not address social characteristics like these.

[2]Hsieh et al. (2018) includes an econometric control for the endogeneity of the collaborations, but their structural model applies to the effort allocation.

These two analyses are strongly complementary, because we focus on the formation of the collaboration network and they focus on the allocation of effort over that network. Separately, our study is one of the first to use a strategic network formation model in an empirical application.[3] By developing a complete empirical application, we demonstrate the extent to which this class of models may be tractable and informative.

However it is also important to note the main limitations in our analysis, and how they affect the economic questions that we can address. The seminal result from Goyal, van der Leij and Moraga-Gonzalez (2006), that the collaboration network in economics is composed of interlinked stars, indicates there is substantial heterogeneity in individual productivity (i.e., which arises from fixed attributes, such as innate "ability"). For computational feasibility, we must abstract from this issue entirely. Empirically, we remove the heterogeneity in individual productivity by residualizing the impact scores for each paper from their authors' average prior impact scores,[4] and then treating researchers as having uniform ability.[5] As a result, we cannot address issues related to heterogeneous productivity, such as assortativity (e.g., Newman, 2003), and its distributional consequences. Additionally, unlike Hsieh et al. (2018), our model does not involve a choice of effort allocation across multiple projects. This implicitly assumes that once the collaborations are established, researchers exogenously

[3]Several papers in the econometric literature on network formation include insightful empirical illustrations, but the emphasis of those papers is methodological. Mele (2020) is a notable example of an empirical application of these methods.

[4]This assumes that the latent ability of researchers is fully captured by the success of their prior publications. The PRT approach does not allow for unobserved heterogeneity as in Graham (2017), for example.

[5]Thus, our model features horizontal differentiation in skills but not vertical differentiation in ability.

allocate a fixed budget of effort among their projects (as in the coauthorship model in Jackson and Wolinsky (1996), for example). Hence we cannot address issues related to shirking and free riding. These limitations are the price we pay for computational feasibility. Throughout the paper we discuss this trade-off between model richness and computational tractability, and how these omissions could affect our results. We also clarify the conditions under which our estimates would be consistent.

The next section develops the model, and it includes a discussion of how the complexity of outcomes arises even with very parsimonious specifications. Section 3 describes the empirical approach, and it explains how we use the production function to control for individual productivity and take other steps to reduce the computational burden. Section 4 describes the data, and Section 5 presents our results.

2 Model

We develop a model of collaborative production that is described in terms of academic research but would apply more broadly to collaborative environments where heterogeneity in skills and the match of skills to tasks are important features (e.g., Lazear (1999), also the literature surveyed by Acemoglu and Autor (2011)). The agents, referred to as researchers or authors, collaborate on research projects. Let $\mathcal{N}$ denote the set of all researchers, and let $\mathcal{T}$ denote the set of all possible research topics (e.g., combinations of JEL codes; see Section 2.1 for further details). Each researcher $i \in \mathcal{N}$ has an area of expertise, or "skill," based on the topics they have studied in the past: $S_i \subseteq \mathcal{T}$. A research project, p, consists of a topic $T_p \in \mathcal{T}$ and a set of researchers $N_p \subset \mathcal{N}$, each with their own skills. Each researcher can participate in up to L projects, and researcher i's set of projects is denoted $P_i \equiv \{p : i \in N_p\}$. The bound on the number of projects (within a given period of time) would arise if, for example,

there are fixed costs of participation in research projects.

The collection of projects across all researchers defines a *collaboration network*, G. Two researchers are linked in this network if they participate in the same project: $ij \in G$ if $i, j \in N_p$ for some project p. Let E_p be the set of all edges that result from project p. The collaboration network is a multigraph (i.e., multiple links are possible between two nodes) because two researchers may collaborate on multiple projects,[6] and it contains self-links ("loops") that represent sole-authored projects. In addition, each edge in G has an attribute, which represents the topic of the project that defines the edge.

A researcher's utility is the sum of the net utilities from each of her projects—i.e., the benefits minus the costs. The benefit of a project comes from the *expected impact* of the resulting publication, denoted Y_p, which is based on the researcher's observations of the current state of the field. The "impact" of a paper is defined based on a measure of journal quality, a journal "impact score." This emphasizes the career impacts of publishing in more prestigious journals (e.g., Heckman and Moktan, 2020) over the scientific impacts of creating more widely cited research (e.g., Angrist et al., 2020).[7] The expected impact is the output of a production function, y, that takes as inputs the topic of the paper (T_p), the number

[6]Often collaboration networks are represented as weighted graphs, where edge weights correspond to the number of common projects between the pair of nodes. However, in order to preserve information about individual projects, we represent the network as a multigraph.

[7]We use a journal impact factor in our utility function because impact factors are often used in tenure and promotion decisions. While citation rates are perhaps a better measure of actual impact on the field, they are more of a public good—the individual researcher can have some difficulty capitalizing on citations, especially in the short run. Journal impact factor is an instant measure of paper quality, whereas citations are slow to accumulate and thus have a much longer-term effect on personal utility.

of authors ($|N_p|$), their skills ($\{S_j\}_{j\in N_p}$), and their connections to other researchers in the network (derived from G).

The cost of project p to researcher i is denoted C_{ip}, which may represent time, effort, and subjective disutility. This is the output of a cost function, c, plus a cost shock, ϵ_{ip}. We assume that the benefit from a project is shared equally among the collaborators, while the costs may be heterogeneous.[8] The crucial details of the specifications of y and c are presented in Section 2.1, after we define the solution concept.

The general expression for researcher i's utility is

$$U_i(G, \mathbf{T}, \mathbf{S}) \equiv \sum_{p\in P_i} (\beta Y_p - C_{ip}) \tag{1}$$

where $\mathbf{T}$ is the vector of topics of all projects and $\mathbf{S}$ is the vector of skills of all researchers in the network, and β is the utility weight on expected impact. Importantly, despite the additive form of (1), this utility function does not rule out spillovers among an individual's projects because the production function and the cost function may depend on attributes of other projects.

We define equilibrium by saying that a collaboration network is *stable* if: (i) no researcher would prefer to drop any of her projects; and (ii) no group of one to k researchers would prefer to add a project on some topic, where k is the maximum number of collaborators. Thus we have the following definition.

Definition 1 (Stability) *A collaboration network G is stable if:*

(i) For all i and $p \in P_i$, $U_i(G, \mathbf{T}, \mathbf{S}) \geq U_i(G\backslash E_p, \mathbf{T}\backslash T_p, \mathbf{S})$; and

(ii) There exists no $\hat{N} \subset \mathcal{N}$ with $|\hat{N}| \leq k$, $\hat{E} = \hat{N} \times \hat{N}$, and $\hat{T} \in \mathcal{T}$ such that

$U_i(G \cup \hat{E}, \mathbf{T} \cup \hat{T}, \mathbf{S}) \geq U_i(G, \mathbf{T}, \mathbf{S})$, $\forall i \in \hat{N}$, *with*

$U_j(G \cup \hat{E}, \mathbf{T} \cup \hat{T}, \mathbf{S}) > U_j(G, \mathbf{T}, \mathbf{S})$ *for some* $j \in \hat{N}$.

[8]Equal credit is the norm in economics, where authors are listed in alphabetical order.

Condition (i) compares the utility from the existing network against that obtained by removing project p. All researchers must weakly prefer maintaining their current projects. Condition (ii) considers adding an additional project among researchers $\hat{N}$ on topic $\hat{T}$. There must be no such project where all of the potential collaborators would have weakly greater utility and one would have strictly greater utility, compared to the existing network.

2.1 Specifications

To manage the computational burden of estimating a strategic network formation model, it is crucial to develop specifications that are informative yet parsimonious. As will be seen in Section 3.2, our model is computationally intensive but feasible, while a substantially richer model would not be tractable. The specifications must therefore contain only the minimal features needed to capture the economic factors in our model, and our model must restrict attention to only certain aspects of behavior. Also, to reduce dimensionality in the data while maintaining salient empirical relationships, we employ machine learning techniques. Below we first describe the construction of two key variables in the model, paper topics and researcher skills, and then explain the specifications of the production function y and the cost function c.

2.1.1 Topics and skills

To define research topics, we use Journal of Economic Literature (JEL) subject codes as a proxy for different ideas within the field of economics. We apply a novel network-based method, developed in Anderson (2017), to characterize the complex relationships among JEL codes and categorize them into a small number of clusters. We start with a *code-to-code network*, in which the nodes are JEL codes, and codes A and B are connected if both A and B are listed on the same paper. The links between codes are weighted to reflect how

often the two ideas are combined (i.e., how often the two codes co-appear on a paper).[9] Appendix figure A1 provides a visualization of the code-to-code network for papers written in 1998–2007.[10] This represents the state of ideas in the field just prior to our analysis period. Subject codes that are close to each other on this network often appear together on a paper, and those that are distant from each other do not. Papers that combine distant codes are presumably making relatively new connections between economic ideas, while papers that contain only nearby codes are deeply embedded in an existing literature.

Then we apply a standard community-finding algorithm (Blondel et al., 2008) to the code-to-code network, which partitions the JEL codes into subgraphs that are highly connected (i.e., have higher density) relative to the entire network. This yields five different clusters of codes, indicated by the colors in appendix figure A1. Interestingly, these clusters roughly correspond to recognizable subject areas within economics. We refer to them as follows:

1. Business (e.g., management, industrial organization, finance)
2. Macroeconomics (e.g., monetary and fiscal policy, international trade)
3. Applied Microeconomics (e.g., labor, health, education, applied econometrics)
4. Local Economies (e.g., agriculture, natural resources, urban, regional)
5. Methods and Theory (e.g., game theory, mathematical methods, pure econometrics)

Thus the cluster analysis reduces the hundreds of JEL subject codes into five topical areas, which are used to characterize paper topics and researcher skills, as described next.

The topic of a paper is defined as the proportion of its JEL codes in each topical area: $T_p \equiv (\rho_{1p}, \rho_{2p}, \rho_{3p}, \rho_{4p}, \rho_{5p})$, where ρ_{ap} is the proportion of codes in area a. Thus a topic is

[9]The weighting scheme uses a one-directional conditional probability, weighting links by $P(A|B)$, where B is the less common of the two codes. This places less common subject codes close to the more common codes that they are co-listed with most often.

[10]All appendix material is in an online supplement.

an element of the 4-dimensional unit simplex. To define skills, researchers are categorized as a specialist in one of the five topical areas, or as a generalist, based on the topics of their papers published in 1998–2007.[11] We take the simple average of the topics of a researcher's papers from this prior period (i.e., an elementwise average of the vectors of proportions), and designate the researcher as a "specialist" in the topical area where the average proportion exceeds 0.5. If there is no area where the average proportion exceeds 0.5, we designate the researcher as a "generalist." Thus each skill corresponds to a subset of $\mathcal{T}$: the subset of all vectors where the relevant dimension exceeds 0.5, or for generalists the subset where no dimension exceeds 0.5. This discretization of the skills is important because, as an observable characteristic of individuals, their cardinality directly affects the computational complexity of the model (see Section 2.2).

2.1.2 Production function

As described earlier, the production function uses attributes of a research team to generate the expected impact of their chosen project—i.e., a prediction of the impact score of the journal where the resulting paper will eventually be published. The function is specified and estimated as a regression tree. This provides crucial dimension reduction of the production inputs, while also yielding better predictions than for example a linear regression. Broadly, a regression tree creates a partition of the joint support of the explanatory variables, where the predicted value of the dependent variable is its average within each element of the partition.[12] This effectively discretizes the many possible combinations of production inputs

[11]The analysis is restricted to "experienced" researchers who published at least two articles in 1998–2007.

[12]Varian (2014) provides a basic introduction to regression trees and related methods for economists.

in our model, into a much small number of sets that have the same predicted impact score within each set. Vectors of inputs in the same element of this partition can therefore be treated as equivalent (and can be represented with the predicted impact score itself).

The production inputs are the topic of the paper, the number of authors, their skills, and measures of their connections to other researchers outside the research team. The topic (T_p, a vector of five proportions) and the number of authors ($|N_p|$, censored at 4) enter directly. The authors' skills ($\{S_j\}_{j \in N_p}$) are transformed to a scalar measure of their collective "skill deficit" relative to the topic. If at least one author is a specialist in each topical area addressed by the paper (i.e., each a where $\rho_{ap} > 0$), then the skill deficit, denoted Z_p^{def}, is zero. Otherwise the deficit equals the sum of the proportions of the paper's JEL codes in topical areas for which the research team has no specialist. This *a priori* specification of the variable is intended to quantify the match of skills to tasks (e.g., Lazear, 1999). Next, for the connections to other researchers, we use two different measures that are designed to capture two possible externalities from links. One is a negative externality from the division of time across multiple projects, as in the coauthorship model in Jackson and Wolinsky (1996). For this we use the total number of projects that the authors are working on concurrently, with each other or with other researchers, denoted Z_p^{prj}. The other is a positive externality from information spillovers (e.g., Jaffe, Trajtenberg and Fogarty, 2000), from researchers outside the team who may provide useful expertise or advice. For this we use the total number of other individuals in the collaboration network who are directly connected to one or more members of the team, which we refer to as the "team degree," denoted Z_p^{deg}. Finally, given these definitions of the input variables, the production function is

$$Y_p = y(T_p, |N_p|, Z_p^{\text{def}}, Z_p^{\text{prj}}, Z_p^{\text{deg}}) \tag{2}$$

where y is specified as a discrete-valued function defined by the regression tree.

2.1.3 Cost functions

The literature has considered many sources of costs that are relevant to scientific research and collaborative production. Our analysis focuses on two broad sources: communication and coordination difficulties in teams, and learning costs from working on new and unfamiliar topics. We use two alternative specifications of the cost function to assess these factors. They must be considered separately in order to limit the number of possible combinations of the cost variables, so as to maintain the tractability of the model (see Sections 2.2 and 3.2).

The first specification pertains to the costs from communication and coordination difficulties that relate to team size and specialization (e.g., Becker and Murphy, 1992; Polzer, Milton and Swann, 2002; Dessein and Santos, 2006; Cremer, Garicano and Prat, 2007). Accordingly the cost of a project is a function of the number of coauthors and the differences in their skills. The number of coauthors, $|N_p|-1$, is used directly.[13] For the skill differences, we use an indicator for whether any two authors on a project are specialists in different areas, denoted X_p^{dif}. We also include an indicator for whether any of the authors is a generalist, denoted X_p^{gen}, because prior research has suggested that generalists can facilitate communication between individuals with different specialities (e.g., Lazear (1999), Cremer, Garicano and Prat (2007)).

This cost function, c_1, is a linear combination of these variables,

$$c_1(|N_p|, X_p^{\text{dif}}, X_p^{\text{gen}}) \equiv \gamma_0 + \gamma_N(|N_p| - 1) + \gamma_S X_p^{\text{dif}} + \gamma_G X_p^{\text{dif}} X_p^{\text{gen}} \tag{3}$$

where $\gamma \equiv (\gamma_0, \gamma_N, \gamma_S, \gamma_G)$ are the cost parameters. The indicator for a generalist only enters with the indicator for skill differences (i.e., $X_p^{\text{dif}} X_p^{\text{gen}}$) because generalists may only reduce costs when there are differences to mediate. The total cost of project p to researcher i is

[13] As the number of authors is censored at 4, the number of coauthors is censored at 3. This implies that the marginal cost of any additional coauthors is zero.

$C_{ip} = c_1(|N_p|, X_p^{\text{dif}}, X_p^{\text{gen}}) + \epsilon_{ip}$, where ϵ_{ip} is an individual-specific cost shock. Researchers are endowed with one cost shock for each of their L possible projects, so the vector of cost shocks for an individual i is $\epsilon_i \equiv (\epsilon_{i1}, \ldots, \epsilon_{iL})$, and the shock assigned to project p (as the l-th project) is $\epsilon_{il(p)}$. The shocks are specified to have standard normal distributions.

The second specification focuses on the costs from working on an unfamiliar topic, as emphasized in the literature on interdisciplinary research.[14] It uses an indicator for whether the topic is unfamiliar or "new" to a researcher, meaning it is not within her current area of expertise: $X_{ip}^{\text{new}} \equiv T_p \notin S_i$. This cost function, c_2, also includes the number of coauthors, because (for this source of costs) coauthors may reduce the cost of producing research, especially on an unfamiliar topic. As before, it is a linear combination:

$$c_2(|N_p|, X_{ip}^{\text{new}}) \equiv \gamma_0 + \gamma_N(|N_p| - 1) + \gamma_T X_{ip}^{\text{new}} \tag{4}$$

Again, the total cost of a project includes a cost shock: $C_{ip} = c_2(|N_p|, X_{ip}^{\text{new}}) + \epsilon_{ip}$.

2.2 Model complexity

Because the computational burden is the most important obstacle for our analysis, we now discuss how the specifications of the functions and variables above affect the complexity of the model. A useful measure for this complexity is the number of distinct outcomes, meaning those which could yield different utility, that are feasible for an arbitrary individual. The specifications above generate a large but ultimately manageable number of such outcomes.

First, the regression tree for the production function yields a partition of the production

[14]For example, Leahey, Beckman and Stanko (2017) finds that interdisciplinary researchers are less productive overall, and attributes this to both cognitive challenges and communication difficulties. Also note that this specification has both observable and unobservable heterogeneity in costs among coauthors.

inputs into just 10 categories of equivalent vectors (see Section 5.1), compared to the thousands of different vectors observed in the data. Second, the parsimonious definitions of the variables in the cost function have just 9 (8) possible combinations, in the first (second) specification.[15] Joining the cost variables with the production function, there are 43 (36) distinct kinds of research projects (i.e., which could yield different net utility, apart from the cost shocks) with the first (second) cost function.[16] Then, the maximum number of projects for an individual is set at three ($L = 3$), which is the 85th percentile among the authors in our data (see Section 4). Consequently, the number of different possible observable outcomes for a given individual with a particular skill is approximately the number of combinations (with replacement) of one, two, or three elements out of 43 (36), which is about 15,000 (9,000).[17] Finally, multiplying this by the number of individual skills (6: 5 areas of specialization plus 1 general skill), gives an upper bound of approximately 90,000 (54,000) different possible observable outcomes for the individuals in our model.

Some of these outcomes can be ignored because they are not observed. However, as we explain in Section 3.2, in order to achieve computational feasibility many observed but

[15]In c_1, the pair $(X_p^{\text{dif}}, X_p^{\text{dif}} X_p^{\text{gen}})$ may take 3 values, (0,0), (1,0), (1,1), when there are 3 or 4 authors, but only 2 values, (0,0), (1,0), when there are 2 authors, and only 1 value, (0,0), when there is 1 author, so $(X_p^{\text{dif}}, X_p^{\text{dif}} X_p^{\text{gen}}, |N_p| - 1)$ has 9 possible combinations. In c_2, the variable X_{ip}^{new} may be either 0 or 1 for any number of authors, so $(X_p^{\text{new}}, |N_p| - 1)$ has 8 possible combinations.

[16]The number of distinct kinds of projects is not 10×9 (or 10×8, respectively) because not all combinations of production inputs and cost variables are possible (e.g., the number of authors must be the same in the production and cost functions).

[17]Not all combinations are possible for all individuals; for example, by definition a specialist cannot be on a project where there is a skill deficit in her own area of expertise.

uncommon outcomes must be combined, which entails a loss of identifying information. By starting with a parsimonious model that generates a relatively manageable number of possible outcomes, we are able to obtain informative results about the utility parameters despite this loss of information.

Without computational constraints, it would be feasible to have a higher bound on the number of projects per individual. We could also allow for heterogeneity in researcher productivity, for example by using the number and impact of prior publications to infer latent individual productivity levels. This would make it possible to analyze the role of highly productive and highly connected "key players" in the collaboration network, and to examine the effects of assortativity based on researcher productivity. Our model is by necessity limited to having researchers be homogeneous in terms of their innate productivity, so that we can address other important questions about team size and specialization, the match of skills to tasks, and certain spillovers across teams.

3 Empirical Approach

To estimate our model, we apply and extend the approach developed in PRT for the identification of utility parameters in strategic models of network formation. The key idea in this approach is to aggregate individuals who occupy structurally similar network positions, called *network types*, that would provide identical utility conditional on an individual's characteristics and tastes.[18] The specification of the model determines the set of relevant network

[18]Burt (1976) provided a seminal contribution on this idea of aggregating individuals who occupy structurally similar nodes in a network. Auerbach (2015) uses this idea to develop a nonparametric regression model for network outcomes, and his approach is based on local subnetworks, as is ours.

types, which is a list of distinct local network structures that may give different utility to an individual. In other words, the network types are the same as the observable outcomes discussed above in Section 2.2. Formally, each network type is a collection of rooted subgraphs that contain all the nodes up to a maximum utility-relevant distance, which is two in our model. Each individual in the network has a network type—i.e., he or she is the root of one of the various possible rooted subgraphs in one of the network types defined by the model.

[FIGURE 1 HERE]

Figure 1 shows two examples of network types from our data: an econometrician "ET" in panel (a), and a theorist "MW" in panel (b). ET has three projects, two with SK and one with FC. Coauthor SK has a third project with two other researchers, while coauthor FC has a sole-authored project (the loop) and a third project with another researcher. ET and SK have the same area of expertise ($S_{ET} = S_{SK}$), and their two projects are in that area. FC is a generalist (with prior publications on a mix of industrial organization and other applied micro topics), and the project with FC is mainly on industrial organization. MW similarly has two projects with one collaborator FP and a third project with another collaborator EC, and those coauthors have additional projects by themselves or with other researchers. MW and her coauthors are all theorists, and her three projects are in that area.

Omitting researcher skills and paper topics, the two network types in figure 1 are very similar. The only difference is that coauthor SK has two collaborators in his project without the root node in panel (a) while coauthor FP has only one collaborator in the analogous project in panel (b). This indicates the possibility of estimating the effects of skills and topics on utility while holding network structure relatively constant. Also, the one difference in network structure between panels (a) and (b) demonstrates variation in the number of coauthors' coauthors given the same number of coauthors' projects. This indicates the possibility of disentangling the possible positive and negative externalities of links that may

arise from sharing information among researchers versus dividing time across projects.

To recover the utility parameters, model predictions of the relative frequencies of the various possible network types are matched to their observed frequencies in the network. This is like using observed choice probabilities to estimate the parameters of a discrete choice model, because, like the alternatives in a discrete choice model, network types determine utility (up to idiosyncratic shocks). The proportions of individuals of each network type, called the *type shares*, can be computed quite easily, even in a network as large as ours (see Appendix A.1). The observed type shares are then compared against predicted type shares from the model, generated with a particular vector of utility parameters. Any vectors of utility parameters that can generate the observed type shares, while satisfying necessary conditions for equilibrium based on Definition 1, are considered to be compatible with the observed network.[19] (This is a partial identification approach, which recovers a set of parameter vectors that are compatible with the observed network.)

The original definition of network types from PRT, based on a simple graph, must be extended to accommodate the network in our application, which is a multigraph with edge attributes. We do this via a parsimonious representation of the network types, which condenses the details of the local network structure into only the information needed to compute utility and assess stability.[20] Under the specifications from Section 2.1, only the topic (T), number of authors ($|N|$), derived production variables (Z), and derived cost variables (X) are needed to evaluate the production and cost functions, which determine the net utility of a project. Any rooted subgraphs that yield the same values of these variables can be treated as equivalent. Therefore, a network type can be represented as a list of these project

[19]Details on our implementation of the method are provided in Appendix A.

[20]The general representation of network types in PRT consists of a local adjacency matrix, and a vector of characteristics of the nodes in the local adjacency matrix.

characteristics for up to $L = 3$ projects, together with the individual's skill, as follows: $(S_i; (y(T_p, |N_p|, Z_p), |N_p|, X_{ip})_{p \in P_i})$.[21] This naturally accommodates a multigraph, and self-links, because the list can contain multiple projects with the same coauthors, and because the project characteristics are well defined for projects with one author. The edge attributes are the topics, which are included in the project characteristics. With this representation of the network types, the additional features of our network are easily incorporated into the approach from PRT, and there is minimal addition to the computational burden. Also this representation works naturally with our definition of stability, which considers changes in utility when projects are added or removed, rather than individual links.

Broadly, compared to using individual links, the computational advantage of this approach is that the number of network types, while large, can be vastly smaller than the number of dyadic links ($n(n-1)/2$, where $n = |\mathcal{N}|$) and the number of (simple) graphs ($2^{n(n-1)/2}$). For example, compared to the roughly 90,000 possible network types in our model, a network with 1,000 individuals would have nearly 500,000 possible links and over $10^{150,000}$ possible graphs. Furthermore, our empirical network has over 30,000 individuals.

The main computational burden in the approach comes from assessing equilibrium conditions when evaluating candidate vectors of utility parameters. The method uses a quadratic programming (QP) problem to determine whether a given parameterization could generate the observed type shares while satisfying necessary conditions based on Definition 1, and the number of variables in the QP problem largely determines the time and memory needed to solve it.[22] We now briefly explain how that number of variables relates to the number of network types—i.e., how the computational burden arises from the complexity of the model.

[21]The representation is even more parsimonious because vectors of production inputs that yield the same expected impact are equivalent, as noted in Section 2.1.2.

[22]See Appendices A.4 and A.6.

The reader is referred to Appendix A.4 for a full presentation that provides formal definitions and shows how the QP problem assesses necessary conditions for equilibrium.

The variables in the QP problem correspond to the elements within a collection of subsets of network types, called "preference classes." Each subset contains those network types that would satisfy the inequality in part (i) of Definition 1, given an individual's skill (S) and cost shocks (ϵ). In equilibrium, an individual must be one of the network types in their preference class. Furthermore, individuals with the same skill and similar cost shocks would have the same preference class, so each of these subsets corresponds to some region in the support of ϵ. The probabilities of these regions can be computed quite tractably (e.g., via simulation), and so the masses of individuals having each preference class can be determined fairly quickly for a given parameterization. Finally, the variables in the QP problem correspond to the elements of each of these subsets. They allocate the masses of individuals among the network types in their respective preference classes, in order to generate predicted type shares.

The overall point is that the number of variables in the QP problem is equal to the sum of the cardinalities of this collection of subsets of network types. The collection is not nearly as large as the power set, because there are dependencies among the network types that could satisfy part (i) of Definition 1 given S and ϵ, but the collection nevertheless grows exponentially with the number of network types. In our final implementation there are around 465 network types or combined categories of types (see Section 3.2), which results in usually having between 100,000 and 500,000 variables in the QP problem, depending on the specific values of the utility parameters (see Appendix A.6).

3.1 Production function and individual productivity

The production function in our model is another extension to the PRT framework. It is important to our analysis for two reasons. First, because the journal impact scores are

observed, it can be estimated separately from the utility parameters, which allows us to include a relatively rich set of production factors in the model. Second, the production function enables us to control for observable heterogeneity in individual productivity in a simple way, by removing the predictive effect of the authors' average prior impact scores on a paper's expected impact.

The separate estimation of the production function requires an exclusion restriction on the cost function and an exogeneity assumption about the production inputs. The exclusion restrictions are apparent in our alternative specifications of the cost function in Section 2.1.3. One version (c_1) excludes the topic of the paper, the other version (c_2) excludes coauthors' skills, and both versions exclude any information about connections to other researchers.[23] The exogeneity assumption pertains to the difference between the realized journal impact scores used to estimate the production function and their expectations conditional on the production inputs. These differences must be mean zero conditional on the observed inputs.[24] So, for example, researchers cannot have prior information about the idiosyncratic potential of different possible projects when they choose what projects to engage in, because that would raise a selection problem for the chosen projects as in a Roy model. Also, researchers cannot systematically allocate effort across projects based on the observed pro-

[23]Arguments can of course be made that these factors may affect costs; for example, it could be that some topics are harder to work on (e.g., take more time, or generate more consternation) than others. Without an exclusion restriction, the identification of the marginal utility of expected impact relies on functional form. The exclusion restrictions are also important for computational feasibility, however, because they make the cost function more parsimonious (see Section 2.2).

[24]For example, let Y_p^o denote the observed journal impact score for paper p. The assumption is that $E[Y_p^o - y(T_p, |N_p|, Z_p) \mid T_p, |N_p|, \{S_j\}_{j \in N_p}, G] = 0$.

duction inputs. The simplest interpretation of these assumptions is that researchers do not have prior information about the idiosyncratic potential of different possible projects when they choose what projects to engage in, and that a fixed budget of effort is divided evenly across a researcher's projects (as in the coauthorship model in Jackson and Wolinsky (1996)).

Next, as discussed in the introduction, our model does not allow for heterogeneity in individual productivity, although we fully acknowledge this is an important factor in academic research. Because we cannot include this heterogeneity for computational reasons, our approach instead is to attempt to remove the variation in individual productivity from the environment. To do this, we first residualize the impact scores for each paper from their expectations conditional on the authors' average prior impact scores, based on their earlier publications from 1998 to 2007, and then use the residualized scores to estimate the production function. This treats individual productivity as an exogenous, fixed attribute (e.g., latent "ability"), which has an additively separable effect on a paper's impact.[25]

The production function for the residualized impact scores has an intuitive interpretation. It shows how research teams under-perform or over-perform their members' average individual productivity, based on complementarities within the team (e.g., the fit between their collective skills and the task at hand) and their connections to the broader collaboration network. The residualized production function is much smaller than would be a "full" specification that includes the authors' average prior impact scores (see Appendix B.3). Furthermore, this avoids the need to include individual productivity as a researcher attribute. This substantially reduces the complexity of the model, making it computationally feasible, although this approach relies on the exogeneity and separability assumptions noted above.

[25]This residualization further assumes that the effect of the authors' fixed productivity is mean independent of the other production inputs and of the difference between the expected and realized impact scores.

3.2 Techniques to reduce computational burden

Even with the parsimonious specifications of the production and cost functions, and the removal of individual productivity, our model is not computationally feasible without further reduction in the number of network types. We use two techniques to restrict the network types used in estimation, although this entails a loss of identifying information, which makes the recovered set a superset of the theoretically identified set.[26]

The first technique is to limit the set of network types to those that are observed in the data. While the model implies roughly 90,000 possible network types (with cost function c_1), only 4,030 appear in the observed network.[27] Restricting to the observed network types does not affect the assessment of part (i) of Definition 1, which applies to existing links, but it does affect the assessment of part (ii), which applies to nonexisting links. A group of researchers could jointly desire to add some project that results in a network type not found in the observed network, which would violate part (ii). Consequently, by ignoring any such violations involving unobserved network types, certain vectors of utility parameters could be accepted as being compatible with the observed network when in fact they are not.[28] The loss of identifying information may be small, however, because there remain hundreds of possibilities for adding new projects based on the observed network types, and the corresponding moments are used to recover a relatively small number of utility parameters.

Just restricting to the observed network types, however, the number of variables in the

[26]This is similar to limiting the moments used for estimation. Appendix A.4 shows that these alterations would not result in excluding any parameter vectors that are contained in the theoretically identified set.

[27]With c_2, the model implies roughly 54,000, and 4,535 appear in the data.

[28]Specifically, this could loosen the lower bounds on the cost parameters.

QP problem is still infeasibly large. A representative vector of utility parameters would generate tens of thousands of different sets of "satisfactory" network types, which results in millions of variables for the QP problem. For tractability on our computing system, however, the maximum number of variables is about 500,000 (see Appendix A.6). Also, over half of the observed network types appear only once (57% with cost function c_1 and 58% with c_2), but these account for less than 10% of the individuals in the network. Consequently, to make the computations feasible (and possibly to reduce noise in the estimated type shares), we collect network types with relatively few individuals into certain "combined" categories and other purely residual categories.

The combined categories collect network types with similar cost variables so that they can provide information that may help bound specific cost parameters. For example, with cost function c_1 there is a category for types that include a paper where the team has members with different skills plus a generalist, and with cost function c_2 there is a category for types that include a sole-authored paper on an unfamiliar topic (see Appendix A.2). Part (i) of Definition 1 can be partially assessed for these combined categories, using the maximum utility of the types in each collection. Hence these combined categories may help tighten the upper bound on certain cost variables, because if the costs are too high then even the maximum utility within a category cannot support the observed type share for that category.

The purely residual categories collect all other rare network types, and provide much more limited information about the utility parameters. There is one residual category for each researcher skill and each number of papers (from 1 to 3). For each residual category, part (i) of Definition 1 is assessed using the maximum possible utility of any network type for a researcher with that skill and that number of papers.

In total, about 20% of the individuals in the network are collected into one of the combined categories, while under 8% go into one of the purely residual categories. This reduces the

number of network types to under 500 (467 with c_1, 462 with c_2), which is computationally tractable. We typically see 1,000 to 10,000 sets of "satisfactory" network types and 100,000 to 500,000 variables in the QP problem. Constructing and solving the QP problem typically takes between one minute and one hour, and requires between 1 and 100 gigabytes (GB) of physical memory. (See Appendix A.6 for further details.)

4 Data

Our data come from the EconLit database of publications, which contains hundreds of economics journals and, importantly for our analysis, includes JEL subject codes in its records. We use journal articles published in 2009 and 2010 to define the research projects that are the equilibrium outcomes of our model. Earlier articles from 1998 to 2007 are used to observe the skill background and prior output of individual researchers, and articles from 2008 are used to include researchers without any current projects.

The individuals who serve as agents in the model are restricted to "experienced" researchers, who published at least two articles from 1998 to 2007, so that their skills and their average prior impact can be inferred with some accuracy.[29] Articles with both experienced and inexperienced researchers are included, but choices are modeled only for experienced researchers.[30] Researchers with publications in 2008, but not in 2009 or 2010, serve as isolates in the current-period network. They are needed to identify the upper bounds on the net utility of projects (i.e., the lower bounds of certain cost parameters).

[29]Appendix B.1 shows that similar distributions of researcher and project characteristics are obtained when different thresholds are used to define experienced researchers.

[30]All authors are included in the number of authors ($|N_p|$), but only experienced authors are used to construct the other variables in the production and cost functions.

[TABLE 1 HERE]

Table 1 provides information on the current and prior publications of the 30,594 experienced researchers in this population. The plurality of them have one article published in the current two-year period (2009–2010), and in the prior ten years (1998–2007) they published almost seven articles on average. About one third of them write on a new or unfamiliar topic in the current period. As discussed in Section 2.2, the limit on the number of current projects is set at three ($L = 3$), which does not affect 85% of these researchers. For the 15% with more than three papers, we use their top three publications in terms of journal impact score to define their network types.[31] One possible interpretation is that projects beyond a researcher's top three (within a given time period) yield zero marginal utility.

[TABLE 2 HERE]

Table 2 presents the distributions of paper topics and researcher skills, based on the topical areas recovered with the cluster analysis described in Section 2.1.1. The first column shows that, across all papers published in 2009 and 2010, roughly one quarter of their JEL codes are in each of the Business, Macro., and Applied Micro. areas. The proportions of codes in Local Economies and Methods/Theory are about 15 percent and 10 percent, respectively. The second column gives the proportions of papers with at least one code in each area, which has the same ranking of these topical areas. The distribution of researcher skills appears in the third column. Roughly one fifth of researchers are a specialist in each of the Business, Macro., and Applied Micro. areas. Fewer are specialists in Local Economies, and under 5 percent are specialists in Methods/Theory. The remaining researchers, about one quarter of the population, do not have a specialized skill and are categorized as generalists. The last column of Table 2 shows the average proportion of JEL codes on papers published in 1998 to

[31]Papers omitted from one researcher's network type may still be used to define another researcher's type.

2007 that are in the researcher's own topical area. Between 75 and 80 percent of the codes on these earlier papers are in the researcher's area of expertise, on average, which indicates the extent of specialization into these topical areas.

[TABLE 3 HERE]

The characteristics of the projects and research teams in 2009 and 2010 are summarized in table 3. About forty percent of the papers have two authors, and most other papers have one or three authors. Less than two percent of papers have more than four authors, so the censoring of the number of authors should not substantially affect our results. Differences in skill backgrounds ($X^{\text{dif}} = 1$) are rare, and only one percent of research teams have a generalist along with these differences ($X^{\text{gen}}X^{\text{dif}} = 1$). However there is substantial diversity in the "skill deficit" measure. The authors have complete expertise in the topic ($Z^{\text{def}} = 0$) for about one third of the papers, and conversely the authors have no expertise in the topic ($Z^{\text{def}} = 1$) for another third of the papers, while the remaining third has a partial deficit (which peaks at 0.5). On average, the authors on these papers are connected to 4.4 other researchers ("team degree") and have 7.2 total projects among them.

The journal impact scores used to estimate the production function come from Kodrzycki and Yu (2006, table 3, results for "within economics impact").[32] They provide scores for 181 economics journals, calculated using articles published in 2003. We assign a score of zero to all other journals in our data. Examples of these impact scores, for the journals used in the construction of figure 1, are shown in appendix table A6. The scores are normalized to range from 0 to 100. Top general interest journals have scores above 25, other general interest journals and top field journals typically have scores between 10 and 20, and other

[32]This follows Ductor et al. (2014), for example, who also use impact scores from Kodrzycki and Yu (2006). The measure is based on a set of axioms (Palacios-Huerta and Volij, 2004), which sets it apart from other common impact scores.

field journals typically have scores below 10. Seventy journals have impact scores below 1, so the assigned score of zero for journals not reported in Kodrzycki and Yu (2006) places them with many other journals at the bottom of the range. The average impact score among the articles used to estimate the model is 3.2 (table 3), which is similar to the average impact score of the authors' prior publications (3.5, from table 1).

5 Results

5.1 Production function

The production function is estimated using the 39,753 articles summarized in table 3. As described in Section 3.1, the impact scores are first residualized from their authors' average prior impact scores, and then the production function is estimated with the residualized scores. Both steps use regression trees and are computed in R.[33]

[FIGURE 2 HERE]

The estimated production function is presented in figure 2. It shows that papers with more authors ("`num_authors`") have higher expected impact (compare terminal nodes 10 vs. 11 and 13 vs. 14, for example). Among papers with more than one author, those on more Business topics ("`bzfin`") have higher impact (nodes 17-19 vs. 10-11 and 13-14) while those on more Local Economies topics ("`agloc`") have lower impact (nodes 13-14 vs. 10-11 and node 19 vs. 17-18). The fit between the authors' skills and the paper's topic ("`skill_deficit`") is also relevant for the expected impact of certain projects (nodes 17 vs. 18). The authors' links in the collaboration network are not strongly predictive, however. Only for sole-authored papers, having collaborators on other projects ("`team_deg`" >

[33]See Appendix A.3 for further details.

1) predicts *lower* impact (nodes 3 vs. 5-6), which likely reflects the allocation of time toward projects with coauthors.[34] Overall, our estimated production function is consistent with prior results from Ductor (2015) and Ductor et al. (2014), which also find that coauthors increase output, while network measures have only modest predictive power, after controlling for authors' prior output.

For comparison, appendix figure A4 shows a larger regression tree that uses the observed (not residualized) impact scores, which includes the authors' average prior impact scores along with the explanatory variables from our production function. The authors' average prior impact is by far the most predictive variable: it determines most of the splits in the first four levels of the tree and accounts for half of the total splits. The effects of the other variables are generally similar to those in our production function. Broadly, this suggests that our procedure of first residualizing the impact scores and then estimating the production function with the residuals is able to capture the most important relationships between our production inputs and the variation in research impact that is not driven purely by the authors' individual productivity. Appendix B.3 discusses how the complexity of the model would increase if we were to include individual productivity in the production function and the individual characteristics. There would be exponentially more network types, and possibly an even greater increase in the number of variables in the QP problem, because of the increased diversity in individual characteristics and observable outcomes. This ultimately shows why we are limited to a model without heterogeneity in individual productivity.

[34] A different variable with the team's total number of distinct projects was designed to capture the division of researcher time, but the two variables may be highly correlated here. Hsieh et al. (2018) similarly find a negative spillover in the cost of effort across a researcher's own projects.

5.2 Utility parameters

Finally, we present the sets of utility parameters that could be compatible with the observed network. The two alternative specifications of the cost function have different parameters and therefore yield different sets. These sets are constructed via a Markov chain Monte Carlo (MCMC) procedure described in Appendix A.5. Intuitively, the procedure draws candidate vectors of utility parameters via a random walk, and assesses each new draw with the QP problem detailed in Appendix A.4. Thousands of vectors are evaluated, and as more vectors are added the contours of the recovered sets become more precise.

[FIGURE 3 HERE]

Figure 3 plots two-dimensional projections of the recovered sets. Two sets are constructed for each specification of the cost function: one using the complete data (black dots), treated as a population, and the other using a 50% random sample (gray dots), which applies the method for statistical inference developed in Appendix C. The results with the sample are included to illustrate how the recovered sets might change if data were available only on a sample of nodes (for example, if it were too costly to collect data on the full population). These plots apply the identifying restriction that $\beta \geq 0$: researchers have (weakly) greater utility from publishing in more prestigious journals.[35] Informative bounds are found for all parameters, except for γ_G ("cG") in the confidence region using the 50% sample, which ranges from -1.80 to 1.64.[36] Also, notably, the values of β ("b") and γ_N ("cN") in the two recovered sets are quite similar, indicating the robustness of these results.

[35]This restriction would also arise naturally in a specification where the scale of the cost shocks is estimated and the parameter on journal impact is normalized to 1 (then $\beta = \sigma^{-1}$, with $\sigma > 0$). Appendix Figure A2 provides analogous plots without this restriction.

[36]A range of this size has over 90% probability in the distribution of the cost shocks, so effect of this factor on behavior would be unknown except for individuals with extremely

[TABLE 4 HERE]

Table 4 reports the minimum and maximum values of the utility parameters in these sets. The sign of the cost per coauthor (γ_N) is identified as negative in both specifications, using our population data. This indicates that any increases in communication and coordination costs from having more coauthors do not offset the cost savings from greater specialization and division of labor. The signs are not identified for the other cost parameters. To understand this, for example with γ_S, relatively few papers have authors with different skills (see table 3), so the model is able to rationalize the low frequencies of network types with such projects, even when their cost is quite high (i.e., at the upper bound of γ_S). The lower bound of γ_S is smaller in magnitude than the upper bound, because there is a much larger number of individuals who could potentially add a project with a coauthor with a different skill, so the cost of such projects cannot be too low (i.e., they cannot be too beneficial).

Finally, with these results, we can briefly consider a counterfactual environment where the benefit from the expected impact of a paper is discounted by the number of authors (i.e., βY_p becomes $\beta Y_p \left|N_p\right|^{-1}$). While it is generally difficult to run exhaustive counterfactuals with partially identified models, we are able to say that our results do not indicate that the distribution of the number of authors would necessarily change in that environment. That is because any of the parameter vectors in these recovered sets could rationalize the observed network, and the counterfactual would effectively reduce the value of β (by different amounts for different sized teams, but still a reduction toward zero). For nearly all of the parameter vectors in the recovered sets, there are also parameter vectors with any smaller value of β, and the same values of all the other parameters. This is seen in figure 3: most points could move horizontally toward zero and remain in the recovered sets. Thus, regardless of which is the true parameter vector, reducing the value of β (by 1/2, 1/3, 1/4, etc.) would

high or low cost shocks.

yield another vector in the recovered set, which can rationalize the current network with the current distribution of the number of authors. That does not guarantee the distribution would stay the same, but it indicates there is an equilibrium such that it could.

6 Conclusion

The utility parameters reported above provide measures of the incentives that rationalize observed patterns of collaboration in economics, under our specification of researcher preferences. Our model captures the preferences of "regular" researchers, who publish three or fewer papers in a two-year period. This leaves the quantification of the incentives and preferences of star researchers as an important subject for future research. Also our approach is computationally intensive, but an equilibrium network formation model like this is necessary given the joint nature of the choices that produce the observed collaborations.

We find that larger research teams tend to produce papers with higher impact, but they do not increase costs of communication and coordination. In addition, we do not find strong evidence that differences in skill backgrounds increase communication costs. Our estimates indicate that adding one more author to a research team reduces individual costs by 0.1 (on a probit scale) at a minimum, and the maximum possible gain in utility from adding a third author, which combines the increase in expected impact with the decrease in costs, exceeds 0.8. As a whole, our results suggest that the trend in economics toward larger research teams with more diverse skill backgrounds will continue.

References

Acemoglu, Daron, and David Autor. 2011. "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics*. Vol. 4, 1043–1171. Elsevier.

Anderson, Katharine A. 2017. "Skill networks and measures of complex human capital." *Proceedings of the National Academy of Sciences*, 114(48): 12720–12724.

Anderson, Katharine A., Matthew Crespi, and Eleanor C. Sayre. 2017. "Linking behavior in the physics education research coauthorship network." *Phys. Rev. Phys. Educ. Res.*, 13: 010121.

Angrist, Josh, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2020. "Inside Job or Deep Impact? Extramural Citations and the Influence of Economic Scholarship." *Journal of Economic Literature*, 58(1): 3–52.

Auerbach, Eric. 2015. "A Nonparametric Network Regression." working paper.

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

Barnett, Andy H, Richard W Ault, and David L Kaserman. 1988. "The rising incidence of co-authorship in economics: Further evidence." *The review of Economics and statistics*, 539–543.

Becker, Gary S, and Kevin M Murphy. 1992. "The division of labor, coordination costs, and knowledge." *The Quarterly Journal of Economics*, 107(4): 1137–1160.

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

Boschini, Anne, and Anna Sjögren. 2007. "Is Team Formation Gender Neutral? Evidence from Coauthorship Patterns." *Journal of Labor Economics*, 25(2): 325–365.

Bosquet, Clément, and Pierre-Philippe Combes. 2013. "Are academics who publish more also more cited? Individual determinants of publication and citation records." *Scientometrics*, 97(3): 831–857.

Burt, Ronald S. 1976. "Positions in Networks." *Social Forces*, 55(1): 93–122.

Burt, Ronald S. 2001. "Structural Holes versus Network Closure as Social Capital." *Network*, 1(May 2000): 31–56.

Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature*, 51(1): 144–61.

Colussi, Tommaso. 2018. "Social ties in academia: A friend is a treasure." *Review of Economics and Statistics*, 100(1): 45–50.

Cremer, Jacques, Luis Garicano, and Andrea Prat. 2007. "Language and the Theory of the Firm." *The Quarterly Journal of Economics*, 122(1): 373–407.

de Paula, Áureo, Seth Richards-Shubik, and Elie Tamer. 2018. "Identifying Preferences in Networks with Bounded Degree." *Econometrica*, 28(1): 263–288.

Dessein, Wouter, and Tano Santos. 2006. "Adaptive organizations." *Journal of Political Economy*, 114(5): 956–995.

Ductor, Lorenzo. 2015. "Does Co-authorship Lead to Higher Academic Productivity?" *Oxford Bulletin of Economics and Statistics*, 77(3): 385–407.

Ductor, Lorenzo, Marcel Fafchamps, Sanjeev Goyal, and Marco J van der Leij. 2014. "Social networks and research output." *Review of Economics and Statistics*, 96(5): 936–948.

Ellison, Glenn. 2013. "How Does the Market Use Citation Data? The Hirsch Index in Economics." *American Economic Journal: Applied Economics*, 5(3): 63–90.

Freeman, Richard B., and Wei Huang. 2015. "Collaborating with People Like Me: Ethnic Coauthorship within the United States." *Journal of Labor Economics*, 33(S1): S289–S318.

Goyal, Sanjeev, Marco J van der Leij, and Jose Luis Moraga-Gonzalez. 2006. "Economics: An Emerging Small World." *Journal of Political Economy*, 114(2): 403–412.

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

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature*, 51(1): 162–72.

Hamilton, Barton H, Jack A Nickerson, and Hideo Owan. 2003. "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation." *Journal of political Economy*, 111(3): 465–497.

Heckman, James J., and Sidharth Moktan. 2020. "Publishing and Promotion in Economics: The Tyranny of the Top Five." *Journal of Economic Literature*, 58(2): 419–70.

Hong, Lu, and Scott E Page. 2001. "Problem Solving by Heterogeneous Agents." *Journal of Economic Theory*, 97(1): 123–163.

Hsieh, Chih-Sheng, Michael D König, Xiaodong Liu, and Christian Zimmermann. 2018. "Superstar Economists: Coauthorship Networks and Research Output." IZA discussion paper 11916.

Jackson, M. O., and A. Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory*, 71(1): 44–74.

Jaffe, Adam B, Manuel Trajtenberg, and Michael S Fogarty. 2000. "Knowledge spillovers and patent citations: Evidence from a survey of inventors." *American Economic Review*, 90(2): 215–218.

Katz, J Sylvan, and Diana Hicks. 1997. "How much is a collaboration worth? A calibrated bibliometric model." *Scientometrics*, 40(3): 541–554.

Kodrzycki, Yolanda K, and Pingkang Yu. 2006. "New approaches to ranking economics journals." *The BE Journal of Economic Analysis & Policy*, 5(1).

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

Lazear, Edward P. 1999. "Globalisation and the market for team-mates." *The Economic Journal*, 109(454): 15–40.

Leahey, Erin, Christine M Beckman, and Taryn L Stanko. 2017. "Prominent but less productive: The impact of interdisciplinarity on scientists research." *Administrative Science Quarterly*, 62(1): 105–139.

McDowell, John M, and Michael Melvin. 1983. "The determinants of co-authorship: An analysis of the economics literature." *The review of economics and statistics*, 155–160.

Mele, Angelo. 2020. "Does School Desegregation Promote Diverse Interactions? An Equilibrium Model of Segregation within Schools." *American Economic Journal: Economic Policy*, 12(2): 228–57.

Moody, James. 2004. "The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999." *American Sociological Review*, 69(2): 213–238.

Newman, Mark E. J. 2003. "Mixing patterns in networks." *Physical review E*, 67(2): 026126.

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

Polzer, Jeffrey T, Laurie P Milton, and William B Swann. 2002. "Capitalizing on diversity: Interpersonal congruence in small work groups." *Administrative Science Quarterly*, 47(2): 296–324.

Rath, Katharina, and Klaus Wohlrabe. 2016. "Recent trends in co-authorship in economics: evidence from RePEc." *Applied Economics Letters*, 23(12): 897–902.

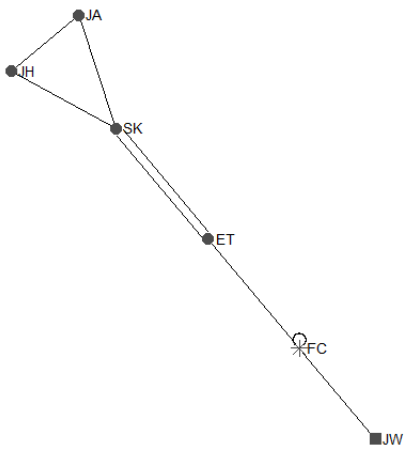
Rhoten, Diana, and Andrew Parker. 2004. "Risks and rewards of an interdisciplinary research path." *Science*, 306(5704): 2046–2046.

Uddin, Shahadat, Liaquat Hossain, and Kim Rasmussen. 2013. "Network effects on scientific collaborations." *PloS one*, 8(2).

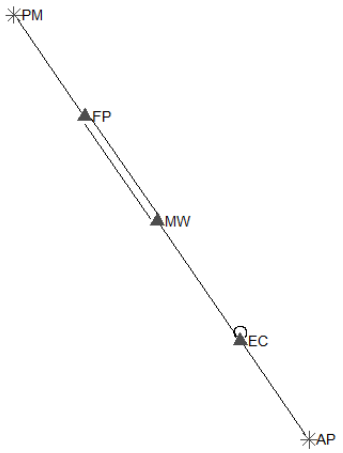
Varian, Hal R. 2014. "Big data: New tricks for econometrics." *Journal of Economic Perspectives*, 28(2): 3–28.

Walsh, John P, and Nancy G Maloney. 2007. "Collaboration structure, communication media, and problems in scientific work teams." *Journal of computer-mediated communication*, 12(2): 712–732.

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



(a) *Researcher ET, econometrician*



(b) *Researcher MW, theorist*

Figure 1: Example network types based on publications in 2009 and 2010

Note: Node shapes indicate areas of expertise, as defined in Section 2.1.1: circle for "applied micro" which also includes much of econometrics, square for "business and finance" which includes industrial organization, triangle for "methods and theory" which includes game theory, and asterisk for a generalist or a new researcher.

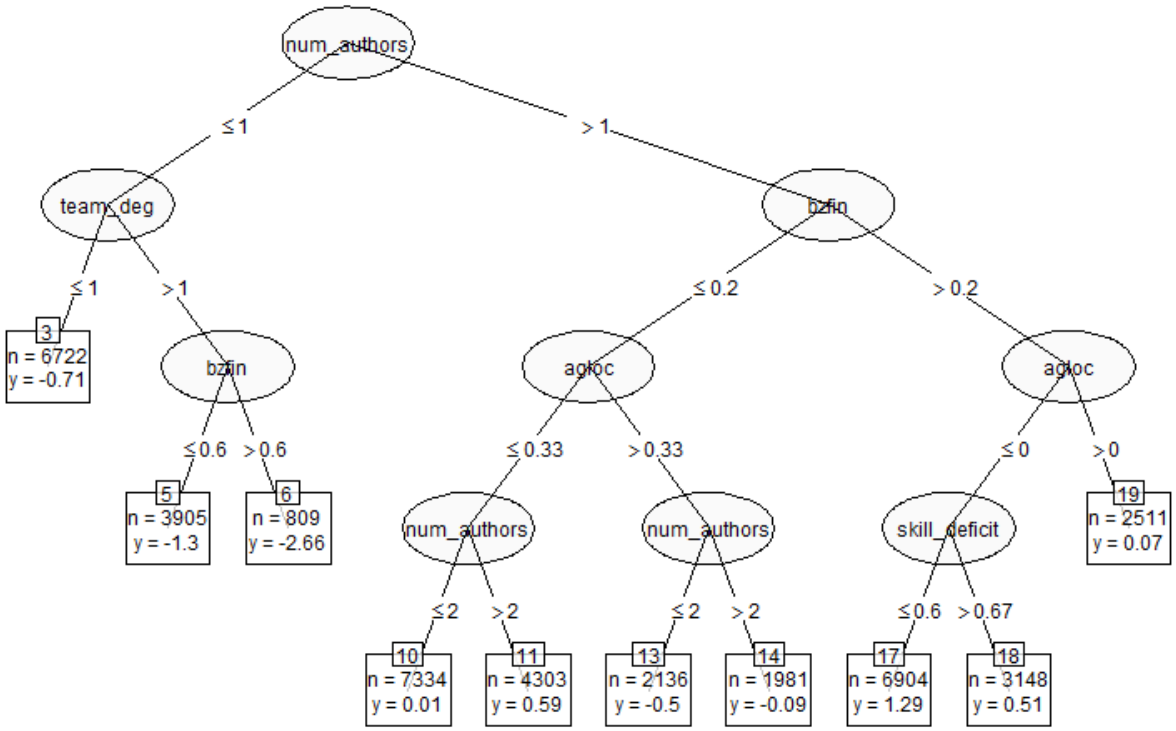


Figure 2: Regression tree estimate of the production function.

Notes: Estimated with the impact scores of articles in 2009 and 2010 that have been residualized from the authors' average prior impact scores from their articles in 1998 to 2007.

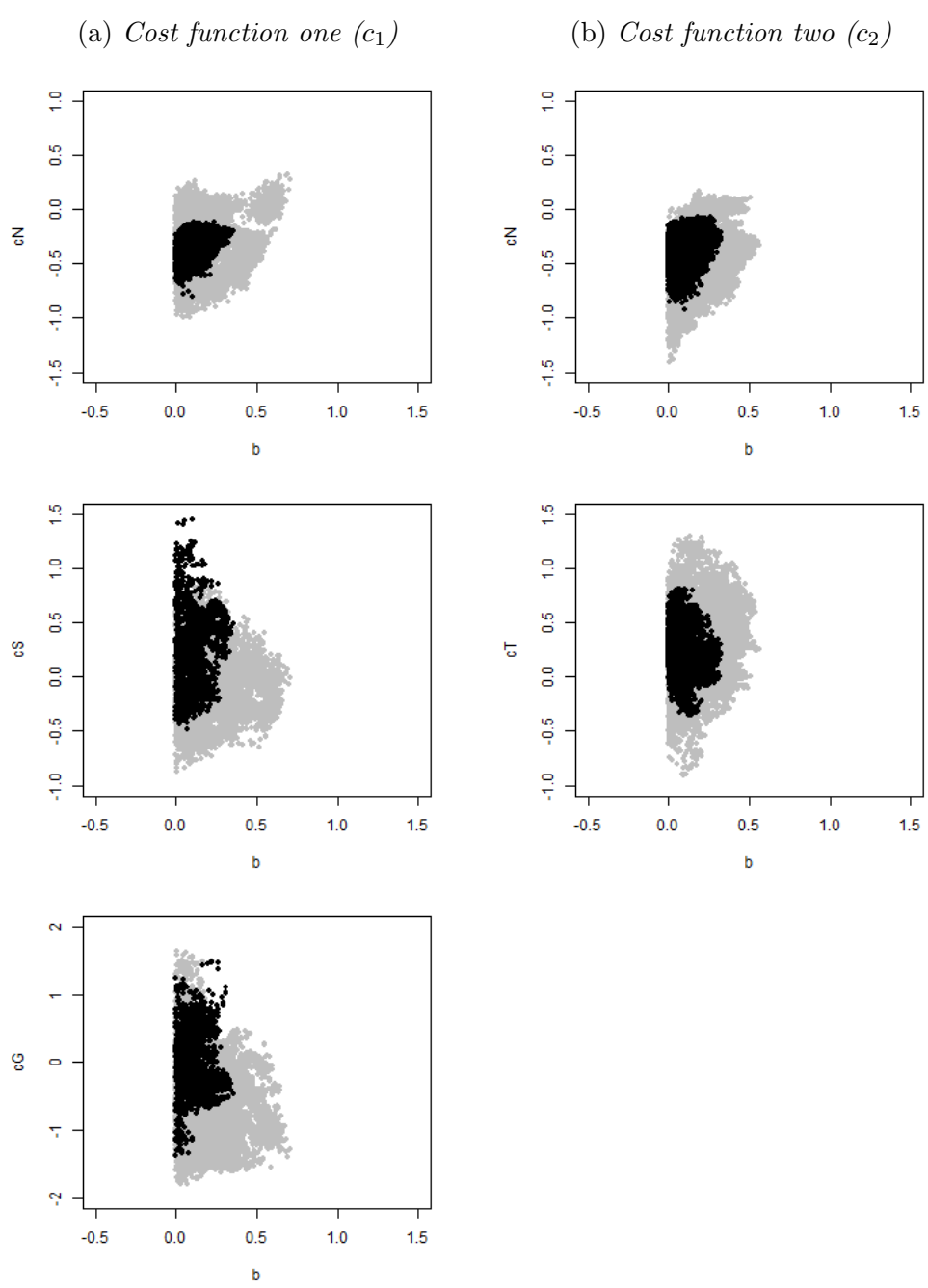


Figure 3: Projections of the recovered sets of utility parameters.

Notes: Axis labels are as follows: b for β, the marginal utility of expected impact; cN, cS, cG, cT for γ_N, γ_S, γ_G, γ_T, the costs of an additional coauthor, of different skill backgrounds, of a generalist given skill differences, and of an unfamiliar topic, respectively. Black dots show vectors recovered with the population data, gray dots with a 50% random sample.

Table 1: Researcher Publications (30,594 individuals*)

Current Publications (2009–2010)		Prior Publications (1998–2007)	
Number of Articles		Number of Articles	
0	0.173	Mean	6.86
1	0.380	SD	6.96
2	0.196		
3	0.106	Average Impact Score	
≥ 4	0.145	Mean	3.48
		SD	7.41
Unfamiliar Topics			
0	0.625		
≥ 1	0.375		

* Researchers who published at least two articles in 1998–2007 and one article in 2008 or 2009–2010.

Table 2: Distributions of Paper Topics and Researcher Skills

	Paper Topics[a]		Researcher Skills[b]	
Topical Area	Dist. of JEL codes (mean prop.)	Prop. of papers w/1+ code in area	Dist. of researcher skills	Prop. of prior codes in own area
Business	0.279	0.439	0.222	0.773
Macro.	0.239	0.377	0.177	0.762
Ap. Micro.	0.231	0.363	0.184	0.797
Loc. Econ.	0.153	0.255	0.118	0.750
Meth./Thy.	0.097	0.166	0.047	0.770
N	39,753 articles		30,594 researchers	

a. Papers published in 2009 and 2010 with at least one experienced researcher (see b) as an author.

b. Researchers who published at least two articles in 1998–2007 and one article in 2008–2010.

Table 3: Project and Team Characteristics (39,753 articles*)

Discrete Measures		Continuous Measures		
	Prop.		Mean	(SD)
Number of Authors		Impact Score	3.20	(9.61)
1	0.288	Residualized Score	0.00	(7.95)
2	0.412			
3	0.229	Team Degree	4.41	(5.37)
4	0.055	Total Projects	7.19	(6.27)
≥ 5	0.016			
Skill Differences		Skill Deficit	0.471	(0.423)
$X^{\text{dif}} = 1$	0.064	$Z^{\text{def}} = 0$ (prop.)	0.354	
$X^{\text{gen}}X^{\text{dif}} = 1$	0.010	$Z^{\text{def}} = 1$ (prop.)	0.323	

* Papers published in 2009 and 2010 with at least one experienced researcher as an author.

Table 4: Recovered Ranges of Utility Parameters

	Cost Function One (c_1)				Cost Function Two (c_2)			
	Population		50% Sample		Population		50% Sample	
Variable (Param.)	Min	Max	Min	Max	Min	Max	Min	Max
Expected impact (β)	0.00	0.35	0.00	0.71	0.00	0.32	0.00	0.56
Cost intercept (γ_0)	1.30	1.96	0.89	2.39	1.20	2.05	0.97	2.51
Num. coauthors (γ_N)	-0.80	-0.12	-1.00	0.32	-0.92	-0.07	-1.41	0.16
Any skill diff's (γ_S)	-0.48	1.45	-0.88	0.88				
Generalist | diff's (γ_G)	-1.38	1.49	-1.80	1.64				
Unfamiliar topic (γ_T)					-0.36	0.81	-0.91	1.30

Notes: Ranges from recovered parameter sets that apply restriction $\beta \geq 0$. Appendix Figure A2 shows the recovered sets without this restriction. Results for 50% sample show ranges from a confidence region that covers the identified set with 95% probability.