Choose Your Adviser Wisely: Endogenous Advisor-Advisee Relationship and Early Stage Coauthorship on Research Output

Jungyoun Kim*
University of Washington

November 14, 2023

Please find the latest version here.

Abstract

We investigate the impact of initial academic social network, formed from advisor-advisee relationships and coauthorships, for economics Ph.D. students (advisees) in the U.S. on their early stage productivity. We define the *academic social network* as a union of i) an advisor-advisee network and ii) a coauthorship network. We model the advisor-advisee relationships with a preferential attachment-like process based on a discrete choice model and find that advisees show weak gender homophilic preferences when choosing advisors. We further model early stage coauthorship formation of advisees through a bipartite network setup, also based on a discrete choice model, and find that advisees prefer to choose projects that are coauthored with their advisors during their graduate studies. Given the *academic social network* through the two networks, we find that the network statistics for advisees have significant positive correlation with early stage output but find weak evidence on the difference by gender. Through simulated synthetic data, we show that, advisee's preference based formation results in productivity gain in percentage at the average individual level but not as much at the aggregate output level, compared to uniform random formation of the networks. This implies that the advisee's preference based allocation of advisors to advisees is less efficient in a social planner's view.

(Results are preliminary and subject to change.)

*kimjy7@uw.edu

1 Introduction

Perhaps the most important first decision a graduate student makes during their program is choosing their advisor. That is, advisors are the closest sources of information, guidance, networking, and collaboration (find citation), which makes them to be the most important asset for a student to acquire before graduation. Numerous work has been done(find citation) in emphasizing the importance of an advisor to their students, both qualitatively and quantitatively, but there has been no attempt on measuring the effect of the allocation process of said asset. This paper takes a novel approach by modeling the advisor-advisee formation process of students in economics Ph.D. programs based in the U.S. through a network growth model. Then we study the effects of those connections on early stage coauthorship formation, and further see how different allocation processes of advisors effect the students at the individual level and the aggregate level.

This study starts with the question of "what if a student met a different advisor?" In order to address this, we first start with the obvious decision process: a student choosing an advisor. During their first couple years in the program, a typical student in a economic PhD program in U.S. based institute chooses their advisor from a pool of choices based on their preference.[1] This forms what is called a genealogy network, namely, and advisor-advisee network. Then, given the relationship, a student starts their early stage research, under some influence by their advisor, if not in a collaborative effort.[2]. If a student engages in a collaboration, then they form a coauthorship network with their coauthor. Lastly, the networks the students formed and the early stage research would further lead to more research output as their academic career expands. Thus, we can answer the question through this channel by modeling each step.

This brings us to our first contribution. We take a novel approach of the advisor-advisee network formation process using the genealogy tree data of the economics literature community members presented by the IDEAS RePEc initiative. Despite advisors playing a key role for the career of an academic as shown above, the literature lacks a quantitive approach on how the process of choosing one works.[3] The two main things to consider in modeling this process is i) advisees have a pool of advisor to choose from (oppotunity set) and ii) the network shows a preferential

[1] We assume a student knows if they will be rejected so the matching is one sided.

[2] Some students may have independent research before meeting their advisor so we consider multiple projects

[3] (Some qualitative approachs)

attachment behavior – i.e., advisors with more students are likely to be more attractive. In order to incorporate these two points, we employ a discrete choice based preferential attachment-like model to formulate the growth process of the network. Namely, we use the number of past students and pairwise attributes as variables in a asymmteric multinomial conditional logit model. The asymmetry here refers to the difference in the advisor opportunity sets of which the students can choose from. We find that the network indeed resembles a preferential attachment process, but also have weak gender homophily.[4]

Next, we model the coauthorship formation process. Similar to how students chose their advisor, students decide to participate in research projects from a pool of perspective projects and for those who choose a collaborative work get to form a coauthorship network with the coauthors. Following Hsieh et al. (2022), we model this coauthorship relationship as a bipartite network but distinct ourselves by modeling the formation process with a discrete choice model. As the genealogy network formation, the asymmetric multinomial conditional logit model allows asymmetric choice sets and under a choice independence assumption, allows for multiple choices as well. We find that, while students are likely to work on a single author project, if they do collaborate, they are likely to work with their advisor or their advisor's coauthor, but not other faculty members (potential advisors to be exact). We also find that with this controlled, we don't observe gender homophily in the early stage coauthorship network.

The last stage allows us to form out second contribution, which is quantifying the difference in allocation of advisors among students – i.e. answering the initial question. In order to do so, we construct a production function that projects the networks statistics of the union network of the two on to a output measurement. Then, we generate synthetic data from the genealogy network formation model, i.e. match different advisors to students. Next, we generate the coauthorship network to match different projects to students, which would be conditional on who their new advisor became. Then using the production function, we calculate the output for each new case.[5] This counterfactual study would allow use to compare the output depending on how the advisors and project are matched. We find that when they are matched according to the model, there are positive percentage gains in average at individual levels compared to a case when advisors and

[4]Gender information is found through *Gender-API.com*. Details are in section 3.1.

[5]These synthetic data generation processes require some strict assumptions. Details are in section 2.4 and section 3.2.2.

projects are matched randomly. However, we find that it is the opposite in the aggregate level, the average total output is greater for random matching. This finding implicates that, decisions made on preferences are dominant strategies for individuals but not an efficient state in the views of a social planner.

Our work shares mainly three branches of the current literature. The first is the area of network formation, a prominent and widely expanding area.[6] We specifically align with the works of Wichmann, Chen and Adamowicz (2016), Overgoor, Benson and Ugander (2020), and Gupta and Porter (2022) which employ a multinomial logit model as a network formation process. Especially for the genealogy network formation, we follow Overgoor, Benson and Ugander (2020) where the connection of discrete choice modeling and preferential attachment is described. Our view on defining the coauthorship network as a bipartite network resembles the work of Hsieh et al. (2022), but the discrete choice modeling of the formation process follows the works of Fu et al. (2017) and Yeung (2019).

The second area is related with coauthorship and collaborative research. Related literature in the field of economics date back to Sauer (1988), but the works of Goyal, van der Leij and Moraga-González (2006) and Azoulay, Zivin and Wang (2010) extend the concept to the network topology, while Fafchamps, Goyal and van der Leij (2010) further studies the formation process of coauthorship networks. Our modeling of the production function borrows the idea of Ductor et al. (2014) in which they find that coauthorship network statistics is useful in predicting future output of a researcher. More recent studies of Ductor, Goyal and Prummer (2021) find how different network characteristics by gender explain the output inequality in research.

The third area are quantified empirical studies on the advisor-advisee relationship[7]. Our work is directly related to the those of García-Suaza, Otero and Winkelmann (2020) and Hilmer and Hilmer (2009), where they find how the quality of the advisors and institutions are positively correlated with the students' early stage performance, especially how students coauthoring with their advisors outperform others, using different datasets. However, we extend the work further to allowing the network related exogenous variables to be endogenous and formulate the process for counterfactual studies.

[6] Graham (2015), Chandrasekhar (2016), de Paula (2017) and de Paula (2020), provide broad reviews in econometrics of network formation.

[7] Qualitative works include

While our scope research does not extensively study homophilic preferences, it is well known that social networks exhibit gender and racial homophily; McPherson, Smith-Lovin and Cook (2001). We choose to use these as control variables in all models thus as a bi-product we observe the strength of homophilc preference in the network formation models as well as gender inequality in research output. Comparable work related to our findings are the likes of Hilmer and Hilmer (2007) and Gaule and Piacentini (2018), which study the impact of advisor-advisee gender matches on research output. We find mixed results compared to the previous literature.

The remaining of the paper has the following structure. We introduce the methodology of the study in section 2, describe the data collection process and definitions of key variables in section 3, report the empirical findings in section 4, and share remarks and conclude in section 5.

2 Methodology

In this section, we first introduce the networks we use in our analysis – the genealogy network and the coauthorship network – and it's corresponding growth processes. Then we define an *academic social network* by combining these two networks which can be seen as human capital a student can accumulate during their studies in graduate school. Given such, we show our empirical strategy on how to measure the impact of the formation processes through a production function of the early stage research of those students. All concepts are illustrated in a simple manner.[8]

Before moving on, we clarify some terminology. The expression *student* and *advisee* is used interchangeably henceforth. Each individual is an *author*, who can have a label of either advisee and/or advisor nor neither. Each paper or working paper, published in a journal or working paper series respectfully, would be addressed as a *project*.

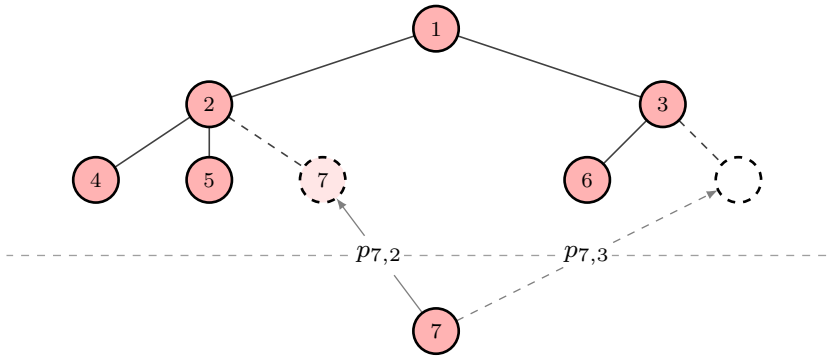
2.1 Genealogy Network and Growth

The genealogy network describes the advisor-advisee relationship. By nature, it is a tree network.[9] The upper part of Figure 1 illustrates an example of a genealogy network. We can see that each node – an author – could be either and advisor (node 1) or an advisee (nodes 4, 5, and 6) or

[8]Detailed mathematical definitions are in Appendix A.

[9]In our dataset, some advisees have two advisors but since the data indicates who is the first and second, we discard the second advisor for our analyses; the number of those how had two advisors was less than 0.5% of the total sample.

Figure 1: Genealogy Network Growth Process – Example



Note: New advisee node 7 faces a pool of potential advisors node 2 or node 3 and decides to choose node 2 with the probability $p_{7,2}$. Since there are only two choices possible, $p_{7,3} = 1 - p_{7,2}$.

both (nodes 2 and 3). The lower part of Figure 1 consists a potential advisee – node 7 – who selects node 2 as their advisor based on a preference structure, from a pool of node 2 and 3 as potential advisors (we assume node 1 is not available here). This preference structure would determine the selection probability for each node, denoted as $p_{7,2}$ and $p_{7,3}$ respectfully in the figure.

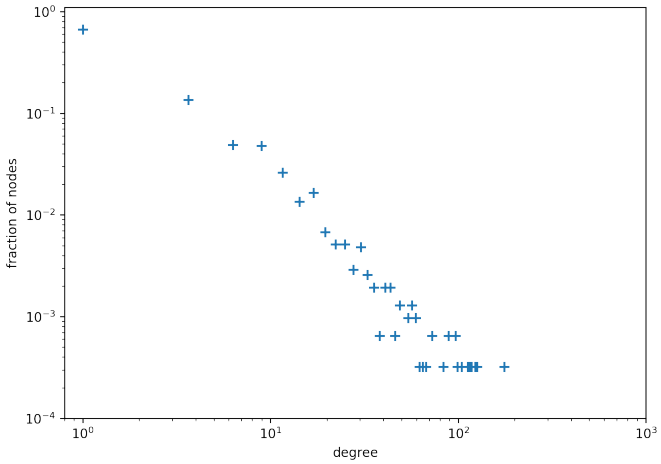
Obtaining these probabilities allows us to formulate the growth process of this network and thus understand how advisee-advisor relationships are formed. We assume that each advisee faces a pool of potential advisors (node 2 and 3 in the example) and would choose an advisor conditional on advisor specific and pairwise attributes. Formally, we define the probability of advisee i choosing advisor j with and asymmetric conditional logit model as

$$p_{i,j} = P(advisee_i = j | j \in AdvisorPool_i) = \frac{\exp(\alpha d_j + z'_{ij}\delta)}{\sum_{k \in AdvisorPool_i} \exp(\alpha d_k + z'_{ik}\delta)}$$

where variable d_j denotes the number of students advisor j has at the time of the selection and z_{ij} denotes a vector of pairwise attribute variables. In our example in Figure 1, we have $AdvisorPool_i = \{2, 3\}$ and thus $d_2 = 2$ and $d_3 = 1$ as the advisor specific attribute. For the pairwise attributes, in case of categorical information such as gender, if node 2 advisor and node 9 advisee are both males, then $z'_{ij} = (1, 0)$.[10]

[10]For categorical data, the dimension of vector z'_{ij} is the number of all categories. For example, if we only have gender data, the dimension would be 2, where each element index is the category for the advisee gender type and the elements are corresponding dummy variables that take value 1 if the gender are the same. So, for a female-female advisor-advisee match, $z'_{ij} = (0, 1)$, and male-female or female-male, $z'_{ij} = (0, 0)$.

Figure 2: Log-Log plot of the unweighted out degree distribution of the genealogy network



Note: The linear slope of the data points suggest that the degree follows a power law / pareto distribution. This is of the out degree of the network since the in degree is always 0 or 1.

This model is an augmented form of a "preferential attachment with fitness" process as described in Overgoor, Benson and Ugander (2020) where we use the number of past students instead of the degree of each advisor and have an asymmetric choice set setup.[11] Our assumption on this approach is based on the nature of the genealogy tree where there are multiple advisee authors connected to one advisor author. Also, the fact that it is more common, at least in the field of economics, for a student to propose to a professor of their choice after observing their characteristics.[12] We also consider the fact that each advisee student has a limited number of advisors to choose from, constrained by both time and place. Figure 2 illustrates the out degree of the genealogy network where the downward sloping linear trend further supports the usage of this approach.[13]

2.2 Early Stage Coauthorship Network Formation

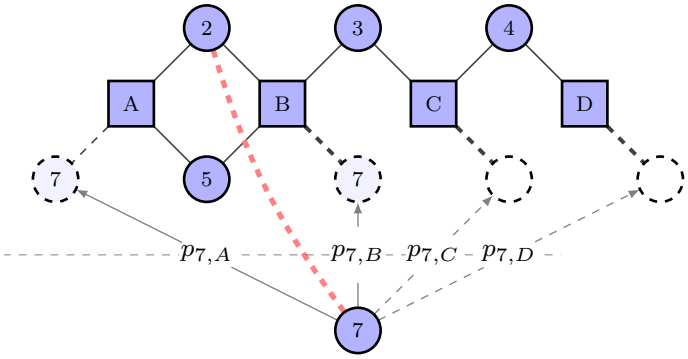
While a coauthorship is a relationship between two authors, the component that connects the authors is a project that they participate in. Thus, to form a coauthorship network, connection,

[11] A preferential attachment model has probability of $p_{i,j} = \frac{d_j^{\alpha}}{\sum_k d_k^{\alpha}}$.

[12] We assume that the advisor author – student – has enough information that they know whether their proposal will be rejected or accepted.

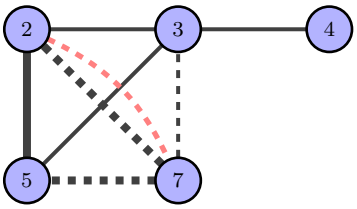
[13] The degree distribution of a network built from a preferential attachment process will have a pareto distriubtion, thus having a down ward sloping linear trend of the log-log plot.

Figure 3: Project Selection Process – Example



Note: Advisee Node 7 is the potential node to be joining the network above. The red dashed line indicates the advisor-advisee relationship between node 7 and 2. With each corresponding probability, node 7 can choose among a pool of projects; A, B, C, and D.

Figure 4: Coauthorship Formation Process – Example continued



Note: Given the choice of node 7, project A and B, projecting the bipartite network on the author set results in the coauthorship network shown above. We can see that the difference in the line width represents the difference in numbers of projects done between coauthors.

an author should be choosing a project, conditional on the information of potential coauthors. Formally, we form the coauthorship network through a author-to-project bipartite network as in Hsieh et al. (2022). Figures 3 and 4 illustrate the process.

In the upper part of Figure 3, we have a bipartite network with 4 authors (nodes 2, 3, 4, and 5) and 4 projects (nodes A, B, C, and D). The projection of this network on to the set of authors would yield the coauthorship network in the upper part of Figure 4, where the width(weight) of the edges are proportionate to the number of projects two authors share. Since author 2 and 5 share two project A and B, the edge connecting the two are thicker (has twice the weight compared to other edges).

Given this configuration, the coauthorship network growth process starts with the advisee author node 7. Recall that advisee node 7 formed an advisor-advisee relationship with author node 2 from the example in Figure 1, which is denoted as the dashed red lines. Conditional on this advisor-advisee relationship and a preference structure, advisee node 7 chooses project A and B with the corresponding probabilities $p_{9,A}$ and $p_{9,B}$ over the set of candidate projects A, B, C, and D. Projecting this network on the set of authors results in the coauthorship network in Figure 4 where we see how the new edges – blacked dashed lines – from advisee node 9 to author nodes 2 and 5 are thicker (twice the weight) than that of the connection to node 3 since both projects A and B involves authors 2 and 5 while author 3 participates in only project B. Thus, if we can formulate the preference structure for the decision process of advisee node 9 choosing projects, we can model the growth process of the coauthorship network.

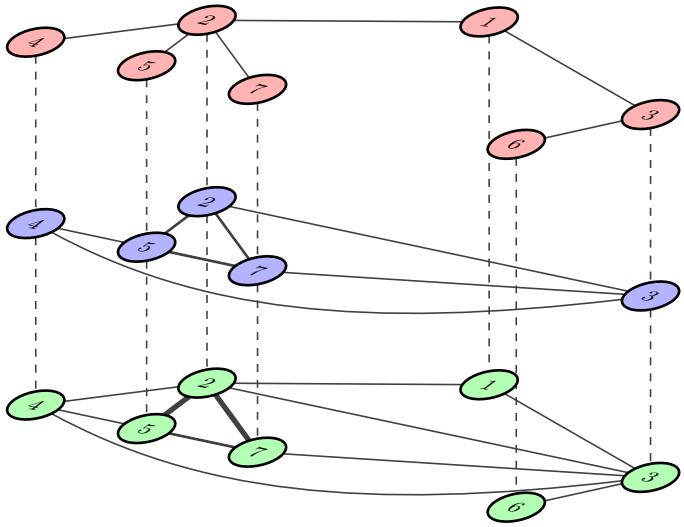
We model the preference structure similar to that of the genealogy network growth process. Formally, with the assumption that each decision is independent, we define the probability of author i choosing project s_n as

$$p_{i,s_n} = P\left(advisee_i = s_n | s_n \in ProjectPool_i\right) = \frac{\exp(q'_{is_n}\theta)}{\sum_{k \in ProjectPool_i} \exp(q'_{ik}\theta)}$$

where vector q'_{is_n} is the vector of pairwise attributes between author i and the coauthors of project s_n.[14] For example, one of the variables is a dummy variable that indicates whether a coauthor of

[14] s_n for $n \in \{1, ..., capacity_i\}$ is the n'th project with a maximum value that takes is the working capacity of advisee i, i.e. the total number of projects advisee i participated in. For example, if advisee i worked on three projects during their graduate studies that $capacity_i = 3$. In the works of Gupta and Porter (2022), they assume the

Figure 5: Academic Social Network – Example



Note: The first and second layer is the genealogy network and coauthorship network respectively. The third layer is the academic social network which can be seen as a union projection of the two networks to the last layer. The edge width does not reflect the depreciation, just the additiveness.

project s is the advisor of author i. In our example above, $q_{9B} = 1$ and $q_{9A} = 1$ but $q_{9C} = 0$ since the coauthor of project A and B includes author node 2, who is author 9's advisor, but not project C. Details on how we construct the variables are described in section 3.2.3 in detail.

2.3 Academic Social Network

We define the *academic social network* using the union of the two social networks; the genealogy and early stage couathorship network. This network can be seen as human capital a student can form during their studies in graduate school to use it for future production: research.

Formally, by having a same additive measurement for the edges of each network, we can simply add one network on top of the other.[15] In order to have the same measurements on each network, we weight each edge by the inverse of the years that have past since the event connecting the two nodes happened. That is, for the genealogy network, the weights would be the inverse of years

independence of choices in cases with mulitple choices in a discrete choice model based network formation. In their setup, the parameters are allowed to vary for each individual (heterogeneous preferences) due to variation introduced by multiple choices. Thus the probability for each individual's decision (of mulitple choices) is the product of the likelihood for each choice. In contrast, we assume homogeneous preferences due to some authors only having one choice observed and thus have a simple likelihood function as here.

[15] We can sum the weighted adjacency matrices to obtain one union network as long as the weights are in same additive measurement.

after graduation, and for the coauthorship network, the weights would be the inverse of the years after publication to a journal or a posting of a working paper. We choose this weighting scheme in the view of considering each publication or the advisor-advisee relationship as an *academic social encounter*, which depreciates over time. It is natural to think that each connection to be less stronger as time goes by, even for advisee-advisor relationships. That is, the initial advisor-advisee connetion has much weight in fresh graduates' academic network, but would gradually decrease unless an advisee frequently cooperate with their advisor throughout their career.

Figure 5 illustrates the concept. Continuing on the previous examples, the new advisee node 7 participates in project A and B conditional on the fact that they chose node 2 as their advisor. The upper and middle layers illustrate this new state, the new genealogy network and new coauthorhip network respectively. Then, we combine the two networks by taking the union of the nodes, all nodes of 1,..., 7, while adding the weights of the corresponding edges.[16] The network in the last bottom layer is the result, which is the academic social network advisee node 7 would be in, by the time of their graduation.

2.4 Empirical Strategy

We aim to measure the impact of the academic social network and its formation process on an advisee's early career performance. To measure the impact, we do so by constructing a log-linear production function as

$$\bar{y}_i = \exp\left(w_i'\gamma + x_i'\beta + \epsilon_i\right)$$

where y_i is the output measure as we define in 3.2.1.

In this function, the parameter of interest is γ which measures the impact of the academic social network as w_i is a vector of the network statistics (human capital). Vector x_i which denotes the control variables, namely, the fixed effects of each advisee such as institution, gender, and region of origin. Note that the time subscript t is not in the right hand side since we try to measure the effect of the initial capital on future production.

Next, to measure the impact of the two network formation processes, we conduct a counterfactual study. Specifically, we use synthetic data generated by each process to obtain the output under

[16] Formal definition is provided in Appendix A.1.

alternative circumstances – different advisor and thus different coauthors – and compare the values to that of what the model predicts with the predicted output from the original data. The synthetic data generation process starts by constructing a genealogy network based on the formation model. For each advisee, we random sample an advisor-author proportionately to the predicted probabilities assigned to each candidate in the pool of advisors. This allows us to construct a new genealogy network, where the original connections are removed and replaced by the newly generated edges. We assume that the graduation year of the advisee does not change, i.e. not dependent on the advisor-author, so the weights on the new edges are calculated accordingly.

Given the new genealogy network, we predict new probabilities using the fitted coauthorship network formation model. In this process, we impose three assumptions that allows us to generate plausible data. The assumptions are as follows.

[1] *The projects and its original authors, sans the advisee, are fixed.* That is, all candidate projects are, in some way, meant-to-happen regardless of who the newly joining coauthor would be which could be seen as a rather strong assumption. However, given the fact that advisees are newcomers to academia while the original authors are mostly likely to be experienced enough to have their on going research pipeline, the formation of the coauthorship could be seen more like a "joining as a branch" from the advisee's prospective rather than a "starting a whole new different project."

[2] *The capacity of each project is fixed.* The number of newly generated joining authors for each project should be the same as the original number of authors. For example, if there were three original authors, two of which are not part of the advisee pool, then only one new advisee can join the project. Similarly, if only one of the original author is not from the advisee pool, then two advisees can jointly join the project. This assumption prevents a project from overwhelming with newly joining advisees.

[3] *The capacity or ability for each advisee is fixed.* As denoted by r_i in section 2.2, the number of projects an advisee joined during their first stage of their career is fixed. This includes the number of solo projects. This assumption prevents the average degree of projects to be contained at a realistic level.

With the assumptions above, for each advisee and their pool of projects, we random sample r_i number of projects without replacement proportionate to the predicted probabilities from the fitted model. During this process, per assumption 2, each advisee selects from an exhaustive pool of projects on a first-come-first-serve basis. If a chosen project has already been taken, then we draw the next random sample with the second highest weight. The order of choosing is shuffled for each iteration of data generation to avoid matching bias.

After collecting all new author-project pairs and constructing a new coauthorship network, we construct the adacemic social network and obtain the corresponding network statistics for each advisee. Then, with the original fixed effects of each individual, we finally collect the corresponding output through the fitted production function. By comparing the output distribution, we gauge the effect of the network formation processes.

3 Data and Variables

3.1 Data Collection

The data is constructed from two sources; the RePEc initiative and *Gender-API.com*.[17] The former assembles a bibiliographic meta database from over 2000 providers relevant to economics including all major publishers and research outlets whereas the latter is an AI powered search which provides services on determining gender and country of origin by name. We collect the necessary data to identify the network structure between the authors and use the names of each author to find the corresponding gender and region of origin. We also construct the output variable based on the publication records for each author.

Starting with the RePEc Genealogy project database, we collect the information of advisees, advisees' advisor, year of graduation, and the institution they graduated from Then, we use the RePEc Author Service which contains information of each author's name and project – published journal or posted working paper – record. Cross-referencing this with the RePEc Publisher data, which includes a list of authors' names and the year of publication/posting for each project, allows

[17] *Gender-API.com* is an online platform that estimates a gender and region of origin based on the first or last name (or both), email, and IP address using AI and machine learning models. Their data sources are publicly available data, governmental data and manual additions/corrections. Ductor, Goyal and Prummer (2021) uses this service to construct their data set, which is used in identifying the productivity difference between male and female authors.

us to construct the time varying coauthorship network.

We collect a total of 59,069 authors and 680,461 projects for the time varying coauthorship network, out of which there are 10,597 authors who are connected in the genealogy network as well. Given the base dataset, we apply a series of filters to select a pool advisees to fit the two network growth models and the production function. First, we select those who graduated between 2006 to 2015 from the department of economics of the top 25% US based institutions as ranked by IDEAS RePEc.[18] Next, we collect those who have at least one effective publication or working paper – project that has a positive output measure – throughout the five years after their graduation and also at least one publication or working paper posted during their graduate studies , i.e. 3 years before graduation or 1 year after to be exact. Finally, we remove the advisees with only one choice in their pool of advisors or pool of projects for each network growth model to ensure identification.

Given the set of the remaining advisees, we identify i) the advisor-authors included in the pool of advisors along with ii) the authors of the projects each advisee can choose for each network growth model, then use *Gender-API.com* to collect the gender and region of origin information.[19] Authors without any gender nor region of origin information are removed from the pool and the filters are applied accordingly. This leaves us with a total of 431 advisees for a pool of 475 advisor-authors and 2,203 projects.

We use the RePEc Publisher database to construct the time varying output variable by collecting the journal and working paper series information. The database includes over 4000 journals and 6000 working paper series, which we select a subsample of 1000 journals and working paper series based on the ranking in CitEc, a RePEc service that provides citation analysis. More detail on how we construct the output variable is described in section 3.2.1.

3.2 Variable Descriptions

3.2.1 Output Measure

We define the time varying output as the research output measure from Ductor, Goyal and Prummer (2021). That is, the sum of the number of publications for the past five years weighted

[18] Ranking as of Sep. 2023 based on all authors and all publication years.

[19] By providing the full name of an author, the API returns a binary result of gender with a probability and a list of possible region of origins with a binary probability of each name coming from each region.

by the quality of the journal or working paper series and discounted by the number of coauthors, for each year. Formally,

$$Y_{it} = \sum_{s=1}^{S_{i,t}} \frac{AIS_s}{(\text{no. of authors})_s}$$

where $S_{i,t}$ is the set of all projects author i published or posted in a working paper series from time t to $t-4$ and AIS_s is the a *article influence score* of the journal or working paper series of project s, which is a measure of quality for said journal or working paper series.

Following Bergstrom, West and Wiseman (2008), we calculate the AIS for journal or working paper series j at time t as

$$AIS_{jt} = \frac{EF_{jt}}{a_{jt}}$$

where EF_{jt} is the *eigenfactor* of journal or working paper series j at year t which solves the following recursive problem

$$EF_{jt} = \sum_{k \in \mathcal{K}} \frac{c_{jk,t}}{\sum_k c_{jk,t}} EF_{kt},$$

and a_{jt} is the normalized project share vector, where each element is the number of all projects in journal or working paper series j divided by the total number of projects in the same sample window collected for time t. Variable $c_{jk,t}$ is the jk-th element in the citation matrix – a 1000 by 1000 matrix given the data set – where each element is the total number of projects in journal or working paper series j in year t that refer to projects published in journal or working paper series k between years $t-1$ to $t-6$; the same sample window to calculate a_{jt}.

Since Y_{it} is extremely right skewed, we take the logarithm of the output plus one to define our final time varying output measure:

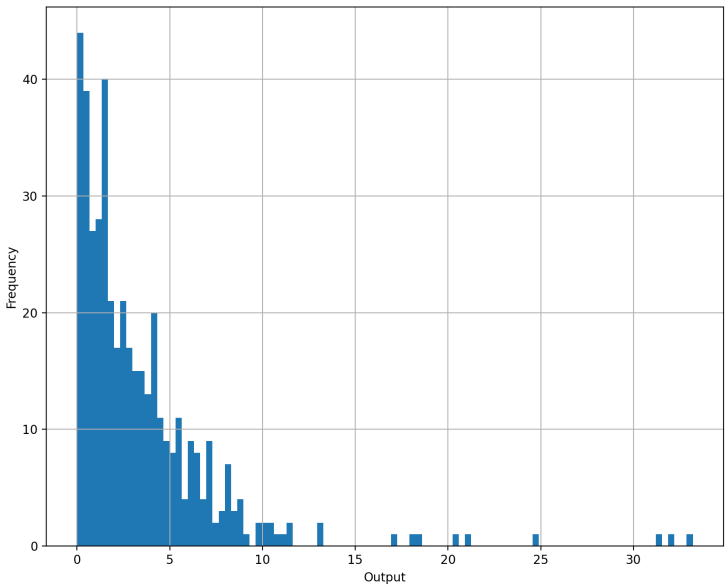
$$y_{it} = \log(Y_{it} + 1).$$

Figure 6 plots the histogram average output level across 6 years for all sample advisees:

$$\bar{y}_i = \frac{1}{6} \sum_{t=0}^{6} y_{i,T_i+t}$$

where T_i denotes the year of graduation. This is the main output measure we use as the dependent variable for the production function and report in section 4. Note that, by definition, $\bar{y}_i$ covers all the output an advisee has produced 5 years before graduation and 5 years after graduation but with

Figure 6: Histogram of the Average Output Level Across 6 Years after graduation – $\bar{y}_i$



Note: Out of 431 observations, even with the log transformation, the output values are exteremly skewed.

most information around the center which is one year after graduation. This allows us to capture the preliminary work done in graduate school but also work done as an independent scholar of an advisee, with an emphasis on the work that is likely most influenced by their advisor.

3.2.2 Pool of Choices – Advisors and Projects

For each network growth model, we allow the advisee to make choices from a pool of choices instead of all possible choices, at least within the data. That is, for the genealogy network, the advisee chooses an advisor-author and for the coauthorship network, they choose a project from the corresponding pool of choices, instead of all advisors and all projects. This assumption is not only realistic to some extent, but is also keeps the model compact and reduce estimation noise.

We define the pool of choices for each case based on the graduation year and institution of each advisee as the following.

[1] *Pool of Advisors.* We assume that an advisee cannot choose an advisor outside of their institution thus we look at all the advisors for advisees from the same institution. Then

Table 1: Summary Statistics for Each Pool to Selection Ratio

	Mean	Median	Variance	Skewness	Kurtosis
Pool of Advisors	0.1759	0.1429	0.1233	1.7458	4.9368
Pool of Projects	0.1463	0.0909	0.1525	2.4811	8.5110

Note: These summary statistics are for the sample of the following ratio for each advisee: the number of true choices over the number of potential choices. For example, each advisee would have a choice of roughly 6 potential advisors to choose from, as per the mean and median. Likewise, assuming the advisee only participates in one project, they have roughly 9 or 10 potential projects to choose from. Compared to the pool of advisors, the pool of projects are much skewed among advisees, mostly due to the difference in research activity among institutions.

for each advisee, we limit the pool of advisors to those who were advisors for the advisees who graduated within the past 5 year window, including their own.[20] The caveat is that we cannot rule out the case of advisors who left or retired from the institution nor those who newly joined but haven't advised any advisee within the 5 year window.

[2] *Pool of Projects.* We define a cohort for each advisee, namely the advisees who graduated the year before, same year, and the year after from the same institution. Then the pool of projects are all the projects of those cohorts, except the ones that they are the sole author of, which publication or posting year is between 3 years before graduation to 1 year post graduation. The idea behind this is that if the cohort of an advisee participated in a project, it is likely that the advisee is capable of doing such as well. The caveat is that the strict compactness doesn't allow any potential outside projects, but given the fact that the time of when these projects are produced is during the advisees' graduate study, restricting the set is not entirely unacceptable.

Table 1 shows the summary statistics of pool to selection ratio, which is the number of choices made of the total size of the pool for each advisee. Thus the for the pool of advisors, it's simply the inverse of the size of the pool, whereas for the pool of projects, it's the number of projects each advisee participated in during the defined window over the total possible projects they could've

[20] Robustness test with 7 years, 10 years show no significant difference in outcome.

Table 2: Number of Advisee, Advisor, and Other Authors Per Gender and Region of Origin

	Advisee	Advisor	Other Authors
No. Obs	431	457	115
Male	338	407	93
Female	93	50	21
Eastern Asia	71	13	16
Eastern Europe	68	50	10
Northern America	148	264	49
Northern Europe	64	128	19
South America	44	22	9
Southern Asia	39	21	15
Southern Europe	135	108	48
Western Asia	57	29	6
Western Europe	112	184	31
Other Regions	65	64	15

Note: Other authors are those that are neither an advisee nor an advisor, but authors for projects in the pool of projects. For the region of origin, each author has up to two categories so the numbers do not add up to the no. of total observations.

participated in.

3.2.3 Fixed Effect Homophily Variables

We collect the gender and region of origin information from *Gender-API.com* and use it to construct variables to account for the homophilic preferences for each network growth model. For each author we search for on *Gender-API.com*, the API returns a JSON file with the information of gender and country of origin. For gender, it returns a binary string value that indicates the gender (*male* or *female*) and the corresponding probability coined with the name that was used to search for. From this, we record the given gender which exceeds probability and remove those samples with *unknown* gender (probability 0.5)[21]

For the country of origin, the API returns a list of countries with positive probability greater than 0.01. A sample result would be in form of $\{USA : 0.84,\ Germany : 0.54,\ Denmark : 0.08\}$,

[21]Only 19 out of 12,635 that we searched for were *unknown*. The average correct probability was 0.962 for male and 0.934 for female.

from which we take the two countries of the highest probability, i.e. *USA* and *Germany*, and use the statistical region – *North America* and *Western Europe* – as defined by Gender-API.com as the region of origin information for each author.[22] Note that results for some names returned only one country so those authors were labeled with one region instead of two.

Table 2 reports the number of each advisee, advisor, and other authors – those who are coauthors of projects in all pool of projects but not advisor-authors – along the corresponding gender and region of origin category. Note that one author can have up to two region of origin categories thus the total count of would not add up to the total number of observations.

Given the labels for each author, in the genealogy network growth model, we measure the degree of homophily by constructing pairwise dummy variables which measures the similarity between te advisee and advisor. To measure gender similarity, we define a dummy variable which takes value 1 if the advisee is the same gender with the advisor-author and 0 if not. Similarly, to measure the region of origin similarity, we define a dummy variable which takes value 1 if the set of region of origin of the advisee shares at least one region of origin (out of the two labels each author has) with the advisor author and 0 if not.

For the coauthorship network growth model, each project can have more than one author, so we don't use a binary variable. Instead, from the group of authors of each project, we collect the categorical information of the rest of the authors who are potential coauthors for each advisee. Then we calculate the ratio of those with the same homophilic characteristics with the advisee. For example, to constuct the variable that measures gender similarity, we calculate proportion of those who have the same gender from the original main authors (excluding her cohort who was the original participating advisee). Thus, for one of a female advisee's potential project which has one male original author and two female original authors, the advisee and project pairwise variable would take value of 2/3. Similarly for the region of origin similarity variable, we calculate the proportion based on how many original authors share at least one region of origin with the advisee. Hence if one out of two original authors share at least one same region of origin, the advisee and project pairwise variable would take value 1/2.

Given the two type of variables: i) advisee advisor-author pairwise variables and ii) advisee

[22] *Gender-API.com* provides three levels of granularity: country, statistical region, continental region. The statistical region is the second level which categorizes 234 countries in 22 groups. Detailed category information can be found at https://gender-api.com/en/api-docs/v2

project pairwise variables, for each case, we construct the advisee fixed effect dummy variables that takes value for 1 for each category and 0 otherwise in each variable. That is, for the gender of advisee, we make a male and female dummy variable separately, eaching taking value of 1 for each corresponding gender and 0 otherwise. Using this, we can construct the interaction term by mulitplying these to similarity variables defined above. Then we can obtain the partial effects of gender similarity for the given gender of the advisee. Similarly for the region of origin, we can obtain the interaction term for each category that the advisee belongs to, though we only use the category with the higher probability of the two.

4 Results

In this section, we present the estimated results and follow up on a counterfactual study based on synthetic data generated from the two fitted network growth models.

4.1 Network Growth Models

4.1.1 Genealogy Network

Table 3 illustrates the estimated results for the genealogy network growth model for which we run three series of regressions through maximum likelihood estimation, where the likelihood is defined as in section 2.1. The reported standard errors in parentheses are based on numerical approximations of the hessian matrix. Regression (1) is the base line, which can be see as the raw preferential attachment model, whereas regression (2) includes the gender homophily variables and regression (3) includes the region of origin homophily variables as well.

The significant estimates on the positive effect of past number of students on the network formation suggest that the genealogy network observes a preferential attachment behavior. Given such, adding the gender and region of origin homophily variables as in model (2) and (3) not changing the estimates much, suggest that even controlling for the homophilic factors, the tendency for new advisees to attach to advisors with more past students is prominent.

For gender homophily, we can see that the estimates are both positive for male and female advisees, suggesting that gender homophilic preference is observed, though it is less significant for the case of female advisees compared to the male advisees. This is mainly due to the lack of

Table 3: Estimated Results for Genealogy Network Growth Model

Variable	(1)	(2)	(3)
no. students	0.0301***	0.0294***	0.0288***
	(0.0050)	(0.0051)	(0.0051)
male–male		0.4307*	0.4335*
		(0.2368)	(0.2376)
female–female		0.4138	0.4263
		(0.3117)	(0.3138)
Eastern Asia			0.2029
			(0.5908)
Eastern Europe			-0.214
			(0.4388)
Northern America			0.4008
			(0.2571)
Northern Europe			-0.2977
			(0.6345)
South America			0.1108
			(0.7256)
Southern Asia			0.2353
			(0.9867)
Southern Europe			-0.0036
			(0.3569)
Western Asia			0.2485
			(0.4495)
Western Europe			-0.0781
			(0.2823)
Other Region			0.9444
			(0.8698)
No. obs	431	431	431
Model Acc.	18.57%	18.70%	18.88%
Rnd. Acc.	17.58%	17.58%	17.58%

*p<0.1; **p<0.05; ***p<0.01

Note: Standard errors in parentheses are based on the approximated hessian matrix from the MLE estimation. Results suggest a clear pattern of preferential attachment from the significancy on the number of students, while weak evidence for gender homophily. Larger standard errors on the female-female coefficient is due to a smaller sample size compared to that of the male-male.

observations of female advisees (total 93) compared to that of male advisees (total 338). On the other hand, none of the region of origin homophily variables show statistical significance, which suggest no evidence of such homophily in the genealogy network.

In order to measure the goodness of fit of the estimation, we calculate the accuracy as the output is categorical. In a typical conditional logit model, if the choice set is homogeneous across all observations, the baseline accuracy is easily calculated – it is simply the probability of choosing one out of the total number of the choice set. However, in this setting, each advisee has a heterogeneous choice set, thus the baseline accuracy criterion is not equivalent for each advisee. Thus we calculate the accuracy of the case where we draw random advisors for each advisee from their corresponding pool as the baseline. The values in Table 3 are calculated based on 10,000 draws each.

In comparing the accuracy values for each regression model, we observe that all three models have at least a higher accuracy compared to the random baseline case, but not by a wide margin. The gap of the accuracy values across each model is narrower, which suggest that the homophilic preferences do not play a strong role in predicting the formation of a new advisor-advisee relationship as the information of the number of past students.

4.1.2 Coauthorship Network

Table 4 shows the estimated results for the coauthorship network growth model. Similar to the genealogy network growth model, we obtain the estimates via maximum likelihood estimation with the likelihood probability defined in section 2.2. Likewise, the reported standard errors in parentheses are obtained by numerical approximations of the hessian matrix. The number of observations for this model is the total number of projects done by all advisees.

We report 4 different regression models, where the baseline – regression (1) – is a model with a single variable; a dummy variable which indicates whether the project advisee participates is a single authored project or not. As we can see, the values are relatively similar across all 4 models which suggests that, unless given pairwise conditions are met, an average advisee would more likely to participate in a single authored paper compared to a coauthored paper.

The rest of the estimated models add the advisor based information, regression (2), or homophiliy preference variable, regression (3), or both – regression (4). The three new variables in model (2) are, for each advisee, i) a dummy variable, which takes value 1 if at least one of the coauthors is their advisor, ii) a dummy variable which takes value 1 if at least one of the coauthors is their advisor's past coauthor, and iii) a dummy variable, which takes value of 1 if the coauthor is another faculty member – this comes from the pool from section 3.2.2. We can see that the advisor

Table 4: Estimated Results for Coauthorship Network Growth Model

Variable	(1)	(2)	(3)	(4)
Not Single Authored	-1.5687^{***}	-1.8379^{***}	-1.600^{***}	-1.7787^{***}
	(0.0819)	(0.0915)	(0.1141)	(0.1185)
Advisor		1.3909^{***}		1.4608^{***}
		(0.1154)		(0.1196)
Advisor's Coauthor		0.6024^{***}		0.5830^{***}
		(0.1228)		(0.131)
Other Faculty		-0.9256^{***}		-0.8535^{***}
		(0.1496)		(0.1526)
Male			−0.0577	-0.1709
			(0.1139)	(0.1221)
Female			0.3632	0.1964
			(0.2362)	(0.2504)
Eastern Asia			0.2028	-0.354
			(0.6131)	(0.7109)
Eastern Europe			0.6226	0.5575
			(0.6589)	(0.7552)
Northern America			-0.3705	-0.5766^{*}
			(0.2987)	(0.3173)
Northern Europe			-0.5545	-0.4209
			(0.7779)	(0.8911)
South America			-0.3519	-0.2175
			(0.8508)	(0.8774)
Southern Asia			1.9006^{***}	1.7758^{***}
			(0.5327)	(0.6338)
Southern Europe			0.7215^{***}	0.9069^{***}
			(0.2980)	(0.3075)
Western Asia			1.1625^{***}	0.5263
			(0.5019)	(0.5363)
Western Europe			0.5549^{***}	0.0614
			(0.2769)	(0.3077)
Other Region			0.9284	0.7924
			(0.6133)	(0.6787)
No. obs	1114	1114	1114	1114
Model Acc.	10.55%	14.80%	11.39%	15.05%
Rnd. Acc.	10.78%	10.78%	10.78%	10.78%

*p<0.1; **p<0.05; ***p<0.01

Note: Standard errors in parentheses are based on the approximated hessian matrix from the MLE estimation. Results suggest advisees prefer to work on a single author project rather than coauthoring, though if they do coauthor, it is likely to be in close proximity with their advisor. Controlling this phenomena, we find no evidence for gender homophily though there are some region of origin homophily observed.

related variables are statistically significant and also positive which suggests that advisees tend to work with their advisors or advisors' coauthors than other faculty members.

Regression (3) tests whether the homophilic preference have significancy for advisees making decisions which we see that, except for several region of origin variables, most are statistically insignificant. Especially, the male gender homophily coefficient is nearly zero and insignificant, suggesting that male advisees have no gender preference when selecting projects. On the other hand, the positive sign and significancy at a 20% level for female gender homophily coefficient suggests weak evidence for female advisees preferring to collaborate with other female coauthors than male coauthors.

Regression (4) includes all variables, where we see the significant estimates of advisor related variables from models (1) and (2) are consistently significant. On the other hand, the estimates on the gender homophily variables overall declined which suggests the likelihood of choosing the alternative choice regarding gender could be partially due to the advisor of the advisee being the same gender.

We report the accuracy of the models in the same manner as in section 4.1.1. We observe that models (2) and (4), namely, the ones with the advisor related variables, have a larger gap of increased accuracy over random selection compared to models without – (1) and (3). Especially, while the difference between model (1) and (3) is less than that of model (2) and (4), which suggest that homophilic preferences have a weak predicting power in the project selections process of advisees.

4.2 Production Function

In this section, we report the results of the production function estimation. We first estimate the mean regression of the log-linear model[23]. Then, we conduct a series of quantile regressions due to our interest in the overall distribution of output. The dependent variable for both regressions is the log of 6 year average of the output measure we defined in section 3.2.1 –$\bar{y}_i$ – including the year of graduation of each advisee.[24] [25]

In table 5, we report 5 models for the mean model, where each model includes institution and

[23]Estimated results for a linear model is in Table B.3. The results do not differ much.
[24]In detail, $\log\left(\frac{1}{6}\sum_{t=0}^{5} y_{i,T_i+t}\right)$.
[25]We conduct a study on each year after graduation up to 5 as well, which the results are in the Appendix.

Table 5: Estimated Mean Regressions Results for the Log-Linear Production Function

	Dependent variable: $\log \bar{y}_i$				
	(1)	(2)	(3)	(4)	(5)
Advisee Male	0.221	0.277	0.165	0.227	0.195
	(0.159)	(0.534)	(0.508)	(0.511)	(0.155)
Advisor Male		0.197	0.126	0.162	
		(0.506)	(0.482)	(0.488)	
Both Male		-0.061	0.027	-0.036	
		(0.568)	(0.541)	(0.542)	
1st order Degree Centrality			2.101***	2.587***	2.611***
			(0.354)	(0.600)	(0.611)
2nd order Degree Centrality				-0.394	-0.413
				(0.467)	(0.474)
constant	1.084*	0.894	0.593	0.524	0.680
	(0.562)	(0.734)	(0.704)	(0.705)	(0.540)
Observations	431	431	431	431	431
R^2	0.294	0.295	0.332	0.333	0.333
Adjusted R^2	0.184	0.181	0.221	0.221	0.224

*p<0.1; **p<0.05; ***p<0.01

Note: Standard errors in parentheses are heteroscedasticity robust standard errors. Centrality measures are from the academic social network. We see clear positive correlation between the 1st order degree centrality network statistics across all models. Extended results are in Table B.2

region of origin fixed effects; we omit to report due to most of the estimates being statistically insignificant.[26] Heteroscedasticity robust standard errors are reported in parentheses.

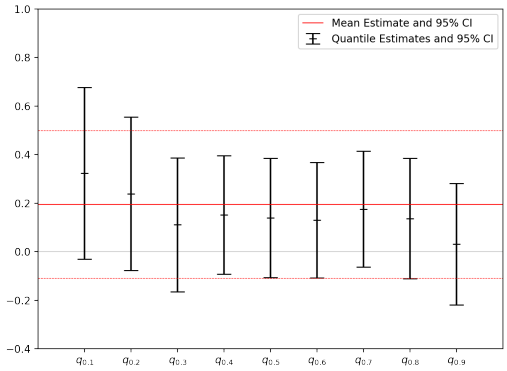
In the first two models, we investigate the effect of advisee and advisor gender on the average output. We find that, in our model, there is no statistical evidence of difference in output gender, nor the cases of advisees having same gender advisors.[27] This contradicts the works of numerous studies such as Ductor, Goyal and Prummer (2021) though this maybe due to our findings focusing on only the early stages of research, while the former finds significance evidence for established researchers throughout their career.[28] Also, contradicting to Gaule and Piacentini (2018), we show that there is no evidence of having an advisor of same gender leading to higher research output, at least in the early stages of research for economics Ph.D. students.

[26]Full results are in the Appendix.

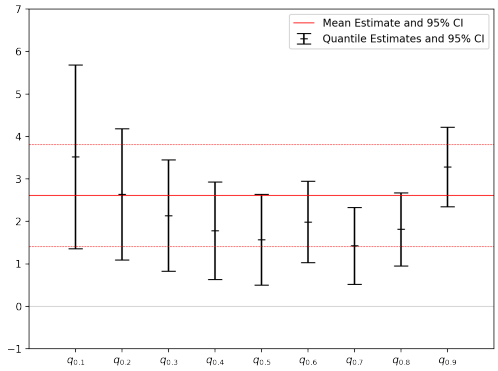
[27]Consistent with Hilmer and Hilmer (2007).

[28]We do find statistically significant difference in a linear model as shown in Table **??** in the Appendix, though it is less significant when controlling for the network statistics.

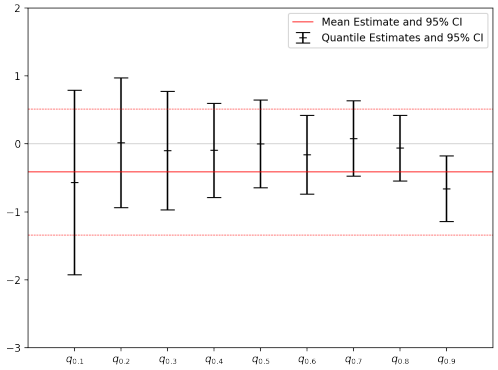
Figure 7: Estimated Quantile Regressions Results for Selected Variables over $\tau \in \{0.1, ..., 0.9\}$



(a) Advisee Male



(b) 1st order Degree Centrality



(c) 2nd order Degree Centrality

Note: Positive correlation between the 1st order degree centrality network statistics and output is observed across all quantiles. No significant difference by gender across all quantiles. Numerical Reports are in Table B.4.

In models (3) and (4), we include the network statistics from the academic social network, namely, the first and second order weighted degree centrality of the each advisee.[29] We find that the 1st order degree centrality is statistically significant as in model (3) and (4), but not the 2nd order degree centrality. Though insignificant, we observe a negative effect of the 2nd order weighted degree on average output.[30] These result suggests that starting with a larger volume of coauthored projects, and consequently coauthors, have a positive impact on early stage research, but also

[29] All network statistic values were multiplied by 10000 for scaling purposes

[30] For higher order network statistics, Ductor, Goyal and Prummer (2021) find that the clustering coefficients are negatively correlated with research output.

Table 6: Network Generation Method for Each Case

	Case 1	Case 2	Case 3
Genealogy	Prediction	Random	Random
Coauthorship	Prediction	Prediction	Random

staying in a relatively smaller network, that is, having connections with less connected authors implies higher productivity.

Figure 7 illustrates the estimates of the coefficient on the 1st and 2nd order degree variable for a series of quantile regressions. Specifically, we use quantile parameter τ to be from 0.1 to 0.9, on the same variables as in model (5) in Table 5 as it has the highest goodness of fit based on the adjusted R^2 values. In both figures, the bar plots plot the estimate and 95% confidence intervals while the solid red line plots the estimates in model (5) and the dashed red lines plots its 95% confidence interval values.[31]

As we see in Figure 7(a), given the control variables, the output difference in gender is mostly statistically insignificant, similar to that found in the mean regression models.[32] We observe similar results in Figure 7(c) for the 2nd order degree as well, where most have statistically insignificant – at 5% level – negative estimates except the coefficient for $\tau = 0.9$. On the contrary, the estimates for the 1st order degree centrality is statistically significant over all quantile levels. Moreover, we see that the estimated outputs for the lower and higher quantiles are more sensitive to the network statistic, compared to those in the middle range.

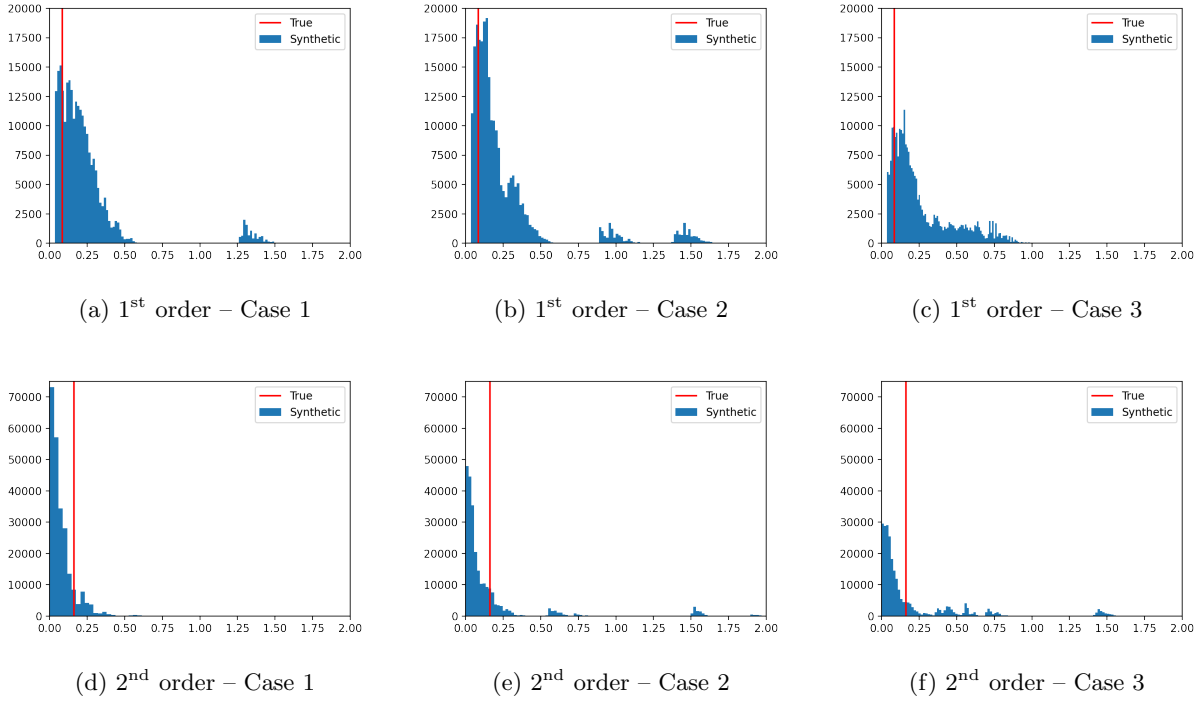
4.3 Counterfactual Study

The counterfactual study aims to find the role of the two network growth models by translating the effect into the output measure. Thus we conduct the study based on synthetic network data generated by each network formation model, then run it through the production function. For each network model, we choose the best performing model in terms of accuracy – model (3) for the genealogy network and model (4) for the coauthorship network – and generate synthetic networks by simulation, following the steps in section 2.4. Namely, we generate 500 genealogy networks and for each simulated genealogy network, we generate 500 synthetic coauthorship networks and

[31] Estimates for all other variables are in the Appendix.

[32] Only the estimate in $\tau = 0.1$ is statistically significant at the 10% level.

Figure 8: Sample of Generated Synthetic Network Statistics – Degree Centrality



(a) 1st order – Case 1 (b) 1st order – Case 2 (c) 1st order – Case 3

(d) 2nd order – Case 1 (e) 2nd order – Case 2 (f) 2nd order – Case 3

Note: Sample from one advisee. Top row: Simulated 1st order degree centrality values for each case. For roughly 78% of the cases, the median value is above the true centrality. Bottom row: Simulated 2nd order degree centrality values for each case. For roughly 40% of the cases, the median value is above the true centrality.

calculate a total of 250,000 predicted output values using model (5) of the production function and the corresponding quantile regression models.

Given the data generation process above, we compare three cases as defined in Table 6. Case 1 is the case described in section 2.4, where advisors are sampled proportionately to the predicted probabilities and given such draws, the projects are sampled proportionately to the predicted conditional probabilities, both over their corresponding pool of choices. Case 2 is where the advisors are sampled randomly, that is proportionate to a uniform distribution over the pool of advisors, and given those samples, the projects are chosen by the conditional probabilities. Case 3 is where both advisors and projects are sampled randomly.

The difference between Case 1 and 2 provides insight on how advisor allocation on advisee effect their predicted productivity. By comparing the output from a random allocation of advisors to a advisee preference based one, we can account how much the advisor selection process takes role in

Table 7: Average of Individual Percentage Gains on Model Predicted Output and Among Cases for Each Case of Synthetic Data based Output

	% Gains on Model Predicted Output			% Gains Among Cases		
	Case 1	Case 2	Case 3	Case 1–2	Case 1–3	Case 2–3
Mean	0.5684	0.5661	0.5550	0.0326	0.0355	0.0303
$q_{0.1}$	0.8714	0.8715	0.8467	0.0575	0.0621	0.0549
$q_{0.2}$	0.6620	0.6487	0.6490	0.0540	0.0572	0.0463
$q_{0.3}$	0.4731	0.4656	0.4638	0.0314	0.0332	0.0261
$q_{0.4}$	0.3734	0.3675	0.3663	0.0223	0.0234	0.0178
$q_{0.5}$	0.3303	0.3235	0.3244	0.0206	0.0213	0.0154
$q_{0.6}$	0.4200	0.4144	0.4116	0.0250	0.0265	0.0208
$q_{0.7}$	0.3050	0.2975	0.2998	0.0202	0.0205	0.0142
$q_{0.8}$	0.3864	0.3796	0.3791	0.0242	0.0253	0.0192
$q_{0.9}$	0.7603	0.7628	0.7394	0.0443	0.0485	0.0435

Left Panel: Mean values of individual percentage gain from the true data predicted output for each model across each case. Roughly 55% gain of the mean model is due to the higher network statistics values than true for all cases. Case 1 is the highest for the Mean model and most quantiles.

Right Panel: Mean values of individual percentage gain from each predicted output of synthetic data cases across each model. Positive values across all rows and columns imply that predicted output based on predicted network formation improves upon random formation everywhere.

predicting an advisee's research output. Case 3 goes a further step, where everything is randomly allocated thus providing a base line for our comparisons.

Figure 8 plots histograms of synthetic network statistics – 1st and 2nd order degree centrality – for each case, that of a sample advisee. The solid red line is the true statistic value for the corresponding advisee. For the synthetic 1st order degree centrality draws, we observe that, for roughly 97% of the 431 advisees, the true value is less than the mean of the generated draws, and for roughly 78%, the true value is less than the median, similarly for all three cases. For the synthetic 2nd order degree centrality draws, the proportion of those with respect to the advisees are 80% and 40%, respectfully.[33]

We first report the results on the individual level gain as in Table 7. The figures in the left panel are the average of individual percentage gain from the true data predicted output to synthetic data predicted output for each case. Note that the gains are around 55% in average, which is due to the

[33]Exact proportions are in Table B.1 of the Appendix.

Table 8: Average Predicted Output ($\hat{\bar{y}}_i$) for Each Model across Each Case

	Case 1	Case 2	Case 3
Mean	3.5510	3.5564	3.5999
$q_{0.1}$	1.3673	1.3702	1.4463
$q_{0.2}$	1.8886	1.9063	1.9255
$q_{0.3}$	2.3413	2.3499	2.3800
$q_{0.4}$	3.0029	2.9900	3.0180
$q_{0.5}$	3.4141	3.3911	3.4169
$q_{0.6}$	4.6166	4.5855	4.6175
$q_{0.7}$	5.4998	5.4735	5.5110
$q_{0.8}$	7.0593	7.0404	7.0647
$q_{0.9}$	12.7217	12.8037	12.8559

Average predicted output for Case 1 is greater than Case 2 for the mid level quantiles while less than the tail quantiles. Note that the predicted aggregate output for Case 3 dominate both cases everywhere. This implies the network formation through the predicted models result in less efficiency in terms of the aggregate productivity.

high proportion of draws of the network statistics being greater than the true values.[34] Therefore, we compare the gains across each case, where we can see that the average gain values for Case 1 are the largest, except for the first and last quantile – even which the difference is negligible compared to Case 3.

For robustness, we also compare the average individual gains among cases, which is reported in the right panel of Table 7. We can see that the first two columns being positive on all models support the findings in the left panel, where Case 1 has the highest average individual gains. Moreoever, unlike how the Case 2 had higher gains to the model predicted output in the first and last quantile, the comparsion between Cases show that Case 1 has gains over both Cases. Thus, for each individual, in average, the predicted output with both model based synthetic data has a gain over the predicted output with synthetic data from either process being uniform random.

Next, we report the results of the predicted output values on the aggregate level in Table 8. Each value is the average of the predicted output for each model across the three cases. To compare Case 1 and 2, the results from the mean regression model show that the average predicted output of Case 1 is smaller, but not as much as a difference there is with respect to Case 3. For the results from

[34]The similar gain amount across all three cases suggests that it is likely due to the low accuracy of the fitted network growth models.

quantile regressions, the average predicted output for Case 2 of the lower quantiles ($q_{0.1}, q_{0.2}, q_{0.3}$) and the highest quantile is larger but smaller for the middle quantiles. This suggests that it is difficult to conclude on whether the advisor allocation based on advisee preferences improves upon random allocation at the aggregate level.

The interesting result is that Case 3 dominates both Cases in all models in terms of average predicted output. This implies that, in the aggergate level, advisee-preference based allocation of advisor and projects are less efficient than that of the case of total random allocation. When comparing this result with the individual level gains, we can see that, in average, each advisee would be better off when they choose their prefered advisor and project, though the overall total research output would be less than that of random allocation.

5 Remarks and Conclusion

As shown in 4, we first find that the genealogy network growth process is mostly a preferential attachment-like formation process where there exists subtle gender homophily between the advisor and advisees. We also find that, advisees are more likely to join projects with their advisor or advisors' coauthor than cases where gender or region of origin are similar. Moreover, we discover that the network statistics from the academic social network are a viable proxy for early stage research output. Also, while there are output difference by gender of advisees, the gender of the advisors and it's match to advisees had no significant explanation power of early stage research output. Through the counterfactual studies, we find some evidence that, compared to random allocation of advisors and advisees, advisee preference based allocation allows advisees to gain more output on the individually. However, we also find that this would result in an overall lower output in average, but also across all predicted quantiles, suggesting that preference based allocation is less efficient than random allocation in the social planner's view.

Our study is an novel attempt on measuring the allocation effect of advisors to students. Our results may shed some light to the areas of higher education and the informatics communitiy, but also policy makers within economics programs.

References

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

Bergstrom, Carl T., Jevin D. West, and Marc A. Wiseman. 2008. "The Eigenfactor™ Metrics: Figure 1." *The Journal of Neuroscience*, 28(45): 11433–11434.

Chandrasekhar, Arun. 2016. *Econometrics of Network Formation.* Oxford University Press.

de Paula, Áureo. 2020. "Econometric models of network formation." *Annu. Rev. Econom.*, 12(1): 775–799.

de Paula, Áureo. 2017. "Econometrics of Network Models." In *Advances in Economics and Econometrics.* , ed. Bo Honore, Ariel Pakes, Monika Piazzesi and Larry Samuelson, 268–323. Cambridge:Cambridge University Press.

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

Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer. 2021. "Gender and Collaboration." *The Review of Economics and Statistics*, 1–40.

Fafchamps, Marcel, Sanjeev Goyal, and Marco J. van der Leij. 2010. "Matching and Network Effects." *Journal of the European Economic Association*, 8(1): 203–231.

Fu, J Sophia, Zhenghui Sha, Yun Huang, Mingxian Wang, Yan Fu, and Wei Chen. 2017. "Two-stage modeling of customer choice preferences in engineering design using bipartite network analysis." Vol. 58127, V02AT03A039, American Society of Mechanical Engineers.

García-Suaza, Andrés, Jesús Otero, and Rainer Winkelmann. 2020. "Predicting early career productivity of PhD economists: Does advisor-match matter?" *Scientometrics*, 122(1): 429–449.

Gaule, Patrick, and Mario Piacentini. 2018. "An advisor like me? Advisor gender and post-graduate careers in science." *Research Policy*, 47(4): 805–813.

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

Graham, Bryan S. 2015. "Methods of identification in social networks." *Annu. Rev. Econom.*, 7(1): 465–485.

Gupta, Harsh, and Mason A Porter. 2022. "Mixed logit models and network formation." *Journal of Complex Networks*, 10(6): cnac045.

Hilmer, Christiana, and Michael Hilmer. 2007. "Women Helping Women, Men Helping Women? Same-Gender Mentoring, Initial Job Placements, and Early Career Publishing Success for Economics PhDs." *American Economic Review*, 97(2): 422–426.

Hilmer, Michael J., and Christiana E. Hilmer. 2009. "Fishes, Ponds, and Productivity: Student-Advisor Matching and Early Career Publishing Success for Economics PhDs." *Economic Inquiry*, 47(2): 290–303.

Hsieh, Chih-Sheng, Michael D Koenig, Xiaodong Liu, and Christian Zimmermann. 2022. "Collaboration in Bipartite Networks." National Taiwan University, Department of Economics Working Papers 2202.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology*, 27(1): 415–444.

Overgoor, Jan, Austin R. Benson, and Johan Ugander. 2020. "Choosing to Grow a Graph: Modeling Network Formation as Discrete Choice." arXiv:1811.05008 [physics].

Sauer, Raymond D. 1988. "Estimates of the Returns to Quality and Coauthorship in Economic Academia." *Journal of Political Economy*, 96(4): 855–866.

Wichmann, Bruno, Minjie Chen, and Wiktor Adamowicz. 2016. "Social networks and choice set formation in discrete choice models." *Econometrics*, 4(4): 42.

Yeung, Fiona Chehong. 2019. "Statistical Revealed Preference Models for Bipartite Networks." University of California, Los Angeles Ph.D. Dissertation.

Supplementary Materials

A Formal Definition of Graphs

A.1 Academic Social Network

Through out the paper, we consider two graphs. The first graph is an author-to-author, time varying unipartite multilayer undirected graph $\mathcal{G}(\mathcal{V}_t, \mathcal{E}_t)$ with an discrete time index t. This graph consists one node set $\mathcal{V}_t$ of size n_t, where each node in this set is considered an author. This set expands by time, in which our setting, could be seen as new Ph.D. graduates joining the community of economics authors. These authors can either be in coauthorship relation, advisors-advisees relation, or both. This is represented in three different types of edge sets: the coauthorship network edge set $\mathcal{E}_t^C$, genealogy network – advisor-advisee relationship – edge set $\mathcal{E}_t^G$, and the academic social network edge set $\mathcal{E}_t^A$. The weights on each element in the edge set, (i, j), respectively, are defined as

$$w_{ij,t}^C = \sum_s \left(\frac{a_{ij,t}^C}{t - T_{ij,s}^P + 1} \right)^{\frac{1}{c}} \qquad \text{for } t \geq T_{ij,s}^P$$

$$w_{ij,t}^G = \left(\frac{a_{ij,t}^G}{t - T_{ij}^G + 1} \right)^{\frac{1}{c}} \qquad \text{for } t \geq T_{ij}^G$$

and

$$w_{ij,t}^A = w_{ij,t}^C + w_{ij,t}^G$$

where for authors i and j, $T_{ij,s}$ denotes the time of encounter of authors for project s, $T_{ij,g}$ denotes the time of encounter of authors as advisor-advisee relationship. [35]. $a_{ij,t}^C$, $a_{ij,t}^G$, and $a_{ij,t}^A$ are elements for corresponding $n_t \times n_t$ adjacency matrices $A_{\mathcal{G}_t}^C$, $A_{\mathcal{G}_t}^G$, and $A_{\mathcal{G}_t}^A$, respectfully. Thus, for authors i and j, if $a_{ij,t}^C = 1$, they had coauthored a project at time t; if $a_{ij,t}^G = 1$, they formed an advisor-advisee relationship at time t; and if $a_{ij,t}^A = 1$, either or both events happened.

We choose this weighting scheme in the view of considering each publication or the advisor-advisee relationship as a *academic social encounter*. It is natural to think that each connection to be less stronger as time goes by, even for advisee-advisor relationships, unless they frequently cooperate for projects. This weighting scheme allows us to i) construct the academic social network by simply adding the weighted adjacency matrix of the two edge sets as they share the same measurement and ii) choose the rate of discount through parameter c, which allows us the flexibility to simply calculate the strength centrality of each author node with a large c if needed.[36] The first property is illustrated in Figure 5 where we can see that the projection of the two first layers of the network to the final layer is the academic social network of all authors in $\mathcal{V}_t$.

[35] We do not observe these actual time so we use the year of publication or presentation for the projects and year of graduation for the empirical work

[36] We set $c = 1$ for simplicity through out the paper.

The second graph is an author-to-project, time varying bipartite undirected graph $\mathcal{H}(\mathcal{V}_t, \mathcal{P}_t, \mathcal{E}_t^P)$ where P_t is the set of projects of size p_t that each author in V_t participates at time t. Let us denote the bi-adjacency matrix of the author-to-project network as $B_{\mathcal{H}_t}$ which has dimensions $n_t \times p_t$ and for each element $b_{is,t}$ in matrix $B_{\mathcal{H}_t}$ takes value 1 if author i participates in project s and 0 otherwise. Then, since

$$a_{ij,t}^C = \mathbf{1}\left\{\left[B_{\mathcal{H}_t} \times B_{\mathcal{H}_t}^{\mathbf{T}} > 0\right]_{i,j}\right\}$$

assigning a weighting scheme for each edge, (i, s) in edge set $\mathcal{E}_t^P$ as

$$w_{is,t}^P = \left(\frac{b_{is,t}}{t - T_{is}}\right)^{\frac{1}{2c}} \qquad \text{for } t > T_{is}$$

allows us to construct the weighted adjacency matrix of the coauthorship network, $[w_{ij,t}^C]_{n_t \times n_t}$, by taking the outer product of the weighted bi-adjacency matrix of the author-to-project network, $[w_{is,t}^P]_{n_t \times p_t}$. Thus, identifying the bi-adjacency matrix $B_{\mathcal{H}_t}$ and the adjacency matrix $A_{\mathcal{G}_t}^G$ with their corresponding weights allows us to fully characterize both graphs.

A.2 Genealogy Network

At time $T_{ij,g}$, Each advisee author i selects an advisor author j from a potential pool of advisor authors $\mathcal{V}_i^{G_a}$ conditional on the pairwise characteristics as well as the advisors' characteristics and advisees' fixed effects. Our assumption is based on the nature of the genealogy tree where there are multiple advisee authors connected to one advisor author and that it is more common, at least in the field of economics, for a student to propose to a professor of their choice after observing their characteristics.[37] We also consider the fact that each advisee student has a limited number of advisors to choose from, constrained by both time and place.

We translate this process to an econometric model by employing an asymmetric conditional logit model which likelihood takes form of

$$P\left(advisee_i = j | j \in \mathcal{V}_i^{G_a}\right) = \frac{\exp(\alpha d_{j,t} + z'_{ij,t}\delta)}{\sum\limits_{k \in \mathcal{V}_i^{G_a}} \exp(\alpha d_{j,t} + z'_{ij,t}\delta)}$$

where the $advisee_i$ is the decision output by advisee-author i. Thus the left hand side denote the probability of advisee author i choosing advisor author j from advisee author i's pool of possible advisors $\mathcal{V}_i^{G_a}$ and establishing an advisor-advisee relationship, i.e. forming an edge in the edge set $\mathcal{E}_t^G$. Note that the advisee chooses only one advisor once, so there the decision is not time varying but the set of pool is, which depends on when the advisee makes the decision.

On the right hand side, the model includes two types of covariates. The first type is the advisors' individual characteristics that are time varying. Namely, denoted as $d_{j,t}$, we use the number of

[37] We assume that the advisor author – student – has enough information that they know whether their proposal will be rejected or accepted.

past students of advisor author j at the time of when advisee author i observes. This is an augmented form of a preferential attachment with fitness process as described in Overgoor, Benson and Ugander (2020) where the degree of the selected node enters the equation as $\alpha \log(d_{j,t})$.[38] Our variation can be seen as the degree minus one (the advisor's advisor), which we leave out since not all of the advisors' advisors are observed, thus could not be calculated. We also neglect the log transformation since there are cases where advisor-authors have no past students. Figure 2 in the main text illustrates the out degree of the genealogy network where the downward sloping linear trend supports the usage of this approach.[39]

The second type are advisor-advisee pairwise fixed effects. Note that the conditional logit model requires covariates to be varying in j so we cannot directly use fixed effects of advisees that only vary in i. However, we can measure the interaction term of a pairwise fixed effect by splitting the datasets into blocks at the cost of efficiency. This is equivalent with multiplying each pairwise fixed effect variable with corresponding dummy variables thus creating a matrix of the dimension of the number of categories. For example, for a given pairwise binary covariate such as gender, we can construct two pairwise covariates of same gender (1 if advisor-advisee have the same gender, 0 if not), one for the male advisees and other for the female advisees and thus measure the difference of preferences by gender.

A.3 Coauthorship Network

As shown in Section A.1, the coauthorship network is fully identified by the weighted bi-adjacency matrix of the author-to-project network, $B_{\mathcal{H}_t}$. In order to model the growth process of this bipartite network, we take a similar approach as the genealogy network case by assuming the advisee authors choose a project from a pool of possible projects they can participated in, conditional on the characteristics of the other participating coauthors. A key difference is that we allow the advisee authors to choose multiple projects regardless of chronological order. We simply assume each advisee-author chooses projects conditional on their genealogy network and pairwise fixed effects from that of with the corresponding coauthors. This assumption relies on the fact that we are only focusing on the first couple projects of the advisee-author, more so on those that they have participated in during graduate school. Thus, the selection process would more likely be effected by the initial academic social encounters; their advisors and advisors' coauthors. This situation also makes it difficult to correctly distinguish the choronological order of projects, hence the myopic setup.

In order to allow the conditional logit model to accompany cases with multiple choices from asymmetric multiple categories, we assume the choices are independent among individuals. Then,

[38] A preferential attachment model in this setup would have a probability of $P(h_i = j | j \in \mathcal{V}_i^{G_a}) = \frac{d_{j,t}^{\alpha}}{\sum_{k \in \mathcal{V}_i^{G_a}} d_{k,t}^{\alpha}}$.

[39] The degree distribution of a network built from a preferential attachment process will have a pareto distriubtion, thus having a down ward sloping linear trend of the log-log plot.

the likelihood of the model takes form of

$$P\left(advisee_i = s_n | s_n \in \mathcal{P}_i, \mathcal{G}\right) = \prod_{s_n}^{r_i} \frac{\exp(q'_{is_n,t}\theta)}{\sum\limits_{k \in \mathcal{P}_{i,t}} \exp(q'_{ik,t}\theta)}$$

where $advisee_i$ is the choice vector, i.e. the decision made by advisee-author i who can choose projects s_n, for $n \in \{1, ..., r_i\}$ where r_i is the number of projects advisee i participates in, from their pool of projects $\mathcal{P}_{i,t}$. Thus the left hand side is the probability of advisee-author i joining r_i number of projects, represented as s, conditional on the pairwise characteristics between the corresponding coauthors represented by $q_{i_n s,t}$. Note that, denoted by r_i, the number of projects each advisee-author i joins could be regarded as the capacity (as referred to in the main text) or ability of the advisee-author and we assume that this not conditional on any information and thus fixed.

B Additional Tables

Table B.1: Proportion of Generated Statistics Greater than True Values

	1st order degree centrality		2nd order degree centrality	
	mean	median	mean	median
Case 1	0.9734	0.7893	0.8111	0.3971
Case 2	0.9709	0.7893	0.8015	0.3656
Case 3	0.9709	0.8111	0.7990	0.3680

Note: These are results show the proportion of where the median or mean of the draws of the networks statistics for each case, exceed the true network statistic value for each advisee sample.

Table B.2: Estimated Mean Regressions Results for the Log-Linear Production Function

	Dependent variable: $\log \bar{y}_i$				
	(1)	(2)	(3)	(4)	(5)
Advisee Male	0.221	0.277	0.165	0.227	0.195
	(0.159)	(0.534)	(0.508)	(0.511)	(0.155)
Advisor Male		0.197	0.126	0.162	
		(0.506)	(0.482)	(0.488)	
Both Male		-0.061	0.027	-0.036	
		(0.568)	(0.541)	(0.542)	
1st order Degree Centrality			2.101***	2.587***	2.611***
			(0.354)	(0.600)	(0.611)
2nd order Degree Centrality				-0.394	-0.413
				(0.467)	(0.474)
constant	1.084*	0.894	0.593	0.524	0.680
	(0.562)	(0.734)	(0.704)	(0.705)	(0.540)
Eastern Africa	-0.167	-0.164	-0.202	-0.244	-0.249
	(0.697)	(0.691)	(0.702)	(0.708)	(0.715)
Eastern Asia	-1.474***	-1.428***	-1.494***	-1.485***	-1.526***
	(0.497)	(0.511)	(0.501)	(0.502)	(0.490)
Eastern Europe	-0.684	-0.669	-0.752	-0.740	-0.755
	(0.497)	(0.501)	(0.486)	(0.486)	(0.483)
Northern Africa	-0.986*	-0.927*	-1.302***	-1.263**	-1.315***
	(0.512)	(0.528)	(0.489)	(0.495)	(0.486)
Northern America	-0.515	-0.493	-0.502	-0.506	-0.527
	(0.474)	(0.479)	(0.470)	(0.471)	(0.467)
Northern Europe	-0.255	-0.244	-0.220	-0.222	-0.232
	(0.556)	(0.562)	(0.566)	(0.567)	(0.562)
South America	-0.663	-0.633	-0.702	-0.697	-0.723
	(0.555)	(0.564)	(0.554)	(0.556)	(0.547)
South-eastern Asia	-0.703	-0.694	-0.553	-0.529	-0.536
	(0.810)	(0.817)	(0.759)	(0.753)	(0.747)
Southern Asia	-0.370	-0.357	-0.432	-0.435	-0.447
	(0.511)	(0.515)	(0.503)	(0.503)	(0.501)
Southern Europe	-0.698	-0.680	-0.720	-0.720	-0.736
	(0.489)	(0.491)	(0.482)	(0.483)	(0.482)
Western Africa	-0.668	-0.604	-0.925	-0.896	-0.953
	(0.939)	(0.967)	(0.640)	(0.644)	(0.619)
Western Asia	-0.490	-0.472	-0.423	-0.417	-0.432
	(0.489)	(0.496)	(0.486)	(0.487)	(0.480)
Western Europe	-0.553	-0.527	-0.534	-0.517	-0.540
	(0.471)	(0.478)	(0.468)	(0.470)	(0.464)
Observations	431	431	431	431	431
R^2	0.294	0.295	0.332	0.333	0.333
Adjusted R^2	0.184	0.181	0.221	0.221	0.224

*p<0.1; **p<0.05; ***p<0.01

Note: Extended table of Table5. Institution fixed effects are omitted. Standard errors in parentheses are heteroscedasticity robust standard errors.[vi]

Table B.3: Estimated Mean Regressions Results for the Log-Linear Production Function

	Dependent variable: $\bar{y}_i$				
	(1)	(2)	(3)	(4)	(5)
Advisee Male	0.834**	0.946	0.491	0.877	0.742*
	(0.402)	(0.945)	(0.837)	(0.847)	(0.383)
Advisor Male		0.409	0.123	0.339	
		(0.913)	(0.839)	(0.844)	
Both Male		-0.123	0.236	-0.151	
		(1.106)	(0.999)	(0.987)	
1st order Degree Centrality			8.539***	11.531***	11.561***
			(2.076)	(2.728)	(2.750)
2nd order Degree Centrality				-2.422**	-2.443**
				(0.959)	(0.991)
constant	1.955	1.559	0.339	-0.086	0.235
	(1.439)	(1.552)	(1.499)	(1.557)	(1.478)
Eastern Africa	1.402	1.408	1.256	0.997	0.993
	(2.263)	(2.248)	(2.292)	(2.330)	(2.344)
Eastern Asia	-1.122	-1.026	-1.294	-1.238	-1.304
	(1.061)	(1.090)	(0.995)	(0.995)	(0.972)
Eastern Europe	-0.158	-0.125	-0.463	-0.392	-0.411
	(1.486)	(1.512)	(1.304)	(1.298)	(1.280)
Northern Africa	-1.068	-0.944	-2.469**	-2.230**	-2.312**
	(1.270)	(1.304)	(1.072)	(1.082)	(1.061)
Northern America	-0.010	0.036	0.002	-0.026	-0.057
	(1.108)	(1.127)	(1.053)	(1.044)	(1.030)
Northern Europe	0.388	0.413	0.508	0.498	0.482
	(1.355)	(1.365)	(1.362)	(1.358)	(1.351)
South America	-0.635	-0.572	-0.852	-0.818	-0.862
	(1.312)	(1.323)	(1.280)	(1.278)	(1.269)
South-eastern Asia	-0.571	-0.553	0.018	0.170	0.162
	(2.109)	(2.130)	(1.965)	(1.952)	(1.936)
Southern Asia	0.198	0.224	-0.081	-0.096	-0.111
	(1.377)	(1.390)	(1.303)	(1.303)	(1.295)
Southern Europe	0.082	0.120	-0.044	-0.040	-0.064
	(1.186)	(1.208)	(1.106)	(1.100)	(1.086)
Western Africa	4.929	5.063	3.758	3.936	3.848
	(5.846)	(5.848)	(4.097)	(4.116)	(4.113)
Western Asia	-0.095	-0.057	0.142	0.182	0.157
	(1.114)	(1.126)	(1.056)	(1.053)	(1.044)
Western Europe	0.238	0.293	0.264	0.368	0.333
	(1.131)	(1.155)	(1.073)	(1.074)	(1.058)
Observations	431	431	431	431	431
R^2	0.294	0.295	0.332	0.333	0.333
Adjusted R^2	0.184	0.181	0.221	0.221	0.224

*p<0.1; **p<0.05; ***p<0.01

Note: Estimated results of the linear production function (non-log-linear). Standard errors in parentheses are heteroscedasticity robust standard errors. Compared to the results of the log-linear model, we can see that the coefficient for the gender dummy variable is statistically significant, but at a 10% level for the model with the best goodness-of-fit. Institution fixed effect estimates are omitted.

Table B.4: Estimated Quantile Regressions Results for the Log-Linear Production Function

	Dependent variable: $\log \bar{y}_i$								
τ	$q_{0.1}$	$q_{0.2}$	$q_{0.3}$	$q_{0.4}$	$q_{0.5}$	$q_{0.6}$	$q_{0.7}$	$q_{0.8}$	$q_{0.9}$
Advisee Male	0.323*	0.239	0.111	0.151	0.139	0.130	0.175	0.137	0.031
	(0.180)	(0.161)	(0.140)	(0.124)	(0.125)	(0.121)	(0.121)	(0.126)	(0.127)
1st order Degree	3.519***	2.636***	2.135***	1.784***	1.569***	1.988***	1.427***	1.813***	3.281***
	(1.101)	(0.787)	(0.667)	(0.583)	(0.542)	(0.487)	(0.460)	(0.438)	(0.476)
2nd order Degree	-0.566	0.018	-0.098	-0.093	0.003	-0.158	0.080	-0.060	-0.658***
	(0.689)	(0.485)	(0.443)	(0.353)	(0.328)	(0.295)	(0.282)	(0.246)	(0.247)
Constant	0.044	-0.572	-0.281	-0.136	0.877	1.591***	1.435***	1.143*	0.504
	(1.057)	(0.995)	(0.749)	(0.621)	(0.606)	(0.591)	(0.543)	(0.587)	(0.763)
Observations	431	431	431	431	431	431	431	431	431
Pseudo R-squared	0.3841	0.2804	0.2264	0.2002	0.1968	0.1895	0.1919	0.2013	0.2466

*p<0.1; **p<0.05; ***p<0.01

Note: Estimate results of quantile regerssions. Standard errors in parentheses are heteroscedasticity robust standard errors. Institution and region of origin fixed effects are omitted.