Série des Documents de Travail

n° 2018-07
Selective matching: gender gap and network formation in research

S.COMBES[1]
P.GIVORD[2]

Les documents de travail ne reflètent pas la position du CREST et n'engagent que leurs auteurs.
Working papers do not reflect the position of CREST but only the views of the authors.

[1] CREST; INSEE. E-mail : stephanie.combes@gmail.com
[2] CREST; INSEE. E-mail : pauline.givord@ensae.fr

Selective matching: gender gap and network formation in research*

Stéphanie Combes and Pauline Givord†

June 16, 2018

Abstract

This paper explores how the academic network extends and its consequences on research outcomes. Using a large academic bibliographic database in research in economics (RePEc), we model first the probability that two researchers collaborate, and secondly the impact of network indicators on the citation rate of research articles. Our results show the existence of a gender-based bias in the researcher matching process. Researchers are more likely to coauthor together when they are of the same gender, even when we control for productivity and proximity in the academic network before they match, as well as unobservable fixed effects of the pair of researcher. This effect is observed mostly at the beginning of the career and fades with the seniority. We also observe that network indicators have a positive impact on the citation index of research articles, suggesting that these selective matching mechanisms may have cumulative effects.

JEL Classification: J24, O31, J45.

Keywords: network analysis; selective matching; gender gap; Probit regression with fixed effects; quantile regression.

*We thank Elian Carsenat, founder of NAMSOR, for his help in data mining and processing, for their research assistant work on a preliminary analysis, and seminar participants at INSEE, JMA, EALE for stimulating discussions. Preliminary research for this paper benefited from the research assistance of Satya Vengathesa, Damien Babet, Julie Djiriguian and Nicolas Paliod and we thank them for their help. We remain solely responsible for the content and possible errors.

†INSEE-CREST. 88 boulevard Verdier 94 Montrouge. Tel: +33 (0)1 87 69 55 65. E-mail: pauline.givord@ensae.fr, stephanie.combes@gmail.com.

1 Introduction

The under-representation of women amongst economists has been documented for a while. Women are missing at each step of the academic ladder, but they are markedly less present at the top ones. They are less likely to get tenured, and it takes them longer to do so (Ginther and Kahn 2004). In 2016, women represent one third of the PhD students in economics but only 15% of full professors, and these figures have hardly changed since the 1990s.[1] Frequent explanations for this continuing situation are that women may have different preferences (Dynan and Rouse 1997) or have a lower productivity for instance because of career breaks due to children.[2]

In this paper, we explore whether different integration in academic networks may provide a complementary explanation to the persistent gender gap in economics. It is indeed commonly assumed that social networks have a key impact on labor outcomes and career prospects. A large professional network, meaning a wide range of formal and informal connections, enhances access to strategic information (for instance hirings or new promotions). Being able to build useful connections within one's own professional environment is therefore expected to improve one's labor prospects. On the other hand, the preexistence of strong social networks, recruiting principally among certain communities and social strata is sometimes accused of being responsible for creating a "glass ceiling" for workers not belonging to these communities. Such phenomenon may be observed if individuals prefer to associate with similar others (Avery et al. 2000). Part of the gender gap may be explained by the fact that women are excluded from the "old-boys network" and thus receive less support for their careers.

Providing empirical evidence on this issue is however difficult, as accurate descriptions of professional networks on a large scale are scarce. Specific surveys provide detailed information on the weak and strong ties of professionals, but they are usually limited to only a few companies or universities. In this regard, the analysis of a large set of academic publications may provide an indirect but useful description of such networks, with homogeneous profiles and outputs.

[1] See the Committee on the Status of Women in Economics Profession's 2016 Annual Report .

[2] Recent contributions, using original data (such as the content of a network forum or the content of c.v.) also suggest that women may be subjected to higher standards than men (Hengel 2017), that they may be less often credited for their contribution to team work (Sarsons Forthcoming), or may be suffered from stereotyping (Wu 2017).

In the first part of the paper, we analyze whether a potential gender-related bias can be identified in the way links are created in the academic network. Such links are identified through co-authorship relations. A collaboration in the writing of a research article is indeed an objective proof of a strong link between two researchers. Using a specification inspired by Fafchamps et al. (2010), we measure whether gender interferes with the decision to coauthor, apart from other observable determinants (previous proximity in the network, respective productivity of both researchers, skill complementarity, etc.) and unobservable components of the pair that may control the expected productive outcome and mutual interests in the match.

In the second part of the paper, we explore the potential consequences of this gender bias, by measuring the network impact in research outcomes. Specifically, we evaluate whether the citation rates received by a researcher for his/her publication are correlated with indicators measuring the integration of this researcher in the academic network, once controlled for individual characteristics. As the distribution of citation is highly skewed, we rely on quantile regressions that provide more accurate measures on the impact of covariates on this outcome.

In practice, the analysis relies on a large bibliometric database extracted from the RePEc (Research Papers in Economics) project. This volunteer-driven initiative offers a very wide coverage of publications in the field of economic research, providing for each article its authors, the year and the journal of publication, as well as the list of the references. The gender of an author is inferred from his/her name and surname.The granularity of the data makes it possible to measure precisely whether a researcher is well-connected within the academic network (and how his/her connections evolve over time) using classical methods in social network analysis (for instance measuring whether an author belongs to a large community of researchers, and whether he/she is central or at the periphery of such communities).

We can also obtain accurate measure of the observed productivity of each researcher since we observe the number of yearly citations received by each article within the database. The PageRank indicator, that can be computed using the entire citation network, provides a measure for research quality and visibility. A high PageRank means that one's publications are not only highly cited, but also highly cited by influential others.

We can estimate the impact of different individual characteristics (and notably gender) on the probability of deciding to work together, controlling

for several observable determinants. We rely on the panel dimension of our database to control for pair-wise specific effects which are stable over time (for instance, having the same native language) using a Mundlack specification. We follow Wooldridge (2016) in order to take into account the fact that our panel is not balanced.

We observe a significant gender-related bias in the process of the network formation. Researchers are more likely to coauthor together when they are of the same gender, even when we control for productivity and proximity in the academic network before they match. Such bias against mixed gender pair is especially marked at the beginning of the career. Though small in magnitude, this impact may have cumulative effects on academic career. Firstly, given that women are under-represented among economists, being less able to connect with men may hamper the integration within the academic network, as well as provide fewer occasions to collaborate on new works. And secondly, because the network integration may have an impact on the indicators used for measuring the research quality. We indeed observe that the number of citations received by an author is positively affected by his/her position in the academic network (as measured by several indicators). Gender differentiation in productivity measures may thus indirectly appear through network effects.

The next section proposes a review of the related literature. The following section presents data and descriptive statistics. In the fourth section we analyze the role of gender in coauthoring decision, and in the fifth the impact of network insertion on the citations process. The final section concludes.

2 Literature Review

This paper relates to two different trends in literature. The first emphasizes the role of networks in the academic market. The positive impact of collaborations on knowledge production is well recognized. Sharing new ideas, benefiting from the experience of others are expected to fuel innovation and research productivity (especially when these others are very productive, (Azoulay et al., 2010)). Well-connected researchers may have greater opportunities to promote their ideas and discuss their researches. They are also more likely to extend further their network. Having a coauthor or colleague in common offers the occasion of being introduced to new people, possibly resulting in new fruitful collaborations. For instance, Combes et al. (2008) observe that "peer-rish" researchers are more likely to graduate the French diploma that conditions tenure position in economics (*Aggrégation du supérieur*). Zinovyeva and Bagues (2015) obtain similar patterns in aca-

demic promotion in Spain. Using random assignation in promotion committees, they observe that candidates strongly benefit from prior connections with committee members. Such patterns may be due to informational asymmetries (committee members may be unable to correctly evaluate the value of works of all candidates because of limited amount of times and imperfect measures of productivity) as well as evaluation biases (preferences for his/her coauthor or former student for instance). Even in situations where no such direct interventions occurs, a good insertion in the researcher network may have a positive spillover on a researcher's career in various ways. Having recognized coauthors can make it easier for a researcher to integrate prestigious institutions (which, in turns, increases the probability of creating fruitful collaborations), as it may constitute a positive signal of future productivity (Ductor et al. 2014). Indeed, quantitative measures of productivity as bibliometric indicators, widely used for many of the appointments, promotions and allocations of research funds, may also be positively affected by a large network. Bosquet and Combes (2013) observe for instance that having several coauthors may also have a positive impact on the citation process. This could be due to the fact that, for instance, a researcher may be invited by one of his/her coauthor to present in seminars or conferences. These meetings may for instance represent opportunities of receiving constructive feedback of one's work, or at least of dissemination of this work. While the research publications volume has sharply increased over the last decades (Card and DellaVigna 2013), being "singled out" is all the more valuable.

The second trend of the literature related to this paper emphasizes the existence of biased preferences in the social network building, and their consequences on professional outcomes in the academic world.

For instance, using data from social networks in U.S. university campuses, Mayer and Puller (2008) show that race is one of the strongest predictors of two students being friend, even when controlling for variety of individual and contextual characteristics. These biased preferences are also observed in professional situations. Considering academic networks, Freeman and Huang (2015) observe that, for US researchers, coauthorship is more frequent with researchers of similar ethnicity. Analyzing the publications of three top economics journals between 1991 and 2002, Boschini and Sjögren (2007) observe that the team formation in economics is not gender neutral. A male researcher is much less likely to coauthor with a woman than a woman does, and surprisingly, that this gender gap in the probability to coauthor with a woman increases with the proportion of female researchers in a subfield (while we would expect the opposite). These results may be mitigated by those of McDowell et al. (2006). Using a rather large sample of economists

(the members of the American Economic Association), they observe that the - effective - differentiated access to the network between men and women seems to attenuate over time, as women become more represented in the profession. As highlighted by Bosquet et al. (2018), female researchers may also be under-represented in the most prestigious institutions partly because of self-censoring mechanisms.

In all cases, most of these analyses focus on specific markets or countries, or limited period of times. They provide very accurate measures, but with limited scopes. Having a very large set of data, which covers a long period of time, allows us to control for a large set of characteristics. Specifically, we are able to control for the coauthors' proximity in the network when a new connection is created, as in Fafchamps et al. (2010). Our approach is very similar to their. However, our emphasis is put on the impact of individual characteristics (and especially gender) on the creation of a link, while they focus on the sole impact of network variables.

3 Data and stylized facts

3.1 Descriptive statistics

We use an extraction from the Repec databases carried out in July 2015. The RePec project is a volunteer-driven initiative launched in the mid-1990s to create a public-access database that promotes wider dissemination of research in economics. The project maintains a database of research papers (IDEAS), articles, books and programs fed by publishers, research centers or directly by the authors (see details in the Appendix A). Listed authors provide contact information and their current affiliations. One may recover personal information such as gender (using here the API Namsor, see details in the Appendix) and proxy of the professional experience (from the date of the first publication, article or working paper, recorded in the database). The entire academic outputs (working papers and academic articles) are also listed.

In order to avoid double counting, we choose to retain only published articles for the main analysis. The dataset includes 318,876 publications, corresponding to 36,822 distinct researchers who have published at least one article in an academic journal. These authors are affiliated to 3,238 distinct research centers.

In this sample, 74% of authors are identified as men and 24% as women

(see Table 1).[3] The gender is obtained using jointly first name, surname and the nationality of affiliation (see details in the Appendix). For 2% of authors in the sample, these information are not sufficient to infer the gender of the researcher with a good accuracy (it corresponds for instance to unisex name as Dominique in France). In the sample, female researchers are on average less experienced, reflecting that the female participation to the labor force is only recent - and especially in the field of research in economics. In our sample, the share of women amongst authors with less than five years of experience is 28% in the late 2000s, while it was only 15% in the early 1990s.

Table 1: Researchers by gender in the Repec Database

	Nb of Obs.	Average experience (years)
All	36,822	14.4
Gender		
Men	27,240	15.3
Women	8,714	11.8
Unknown	868	11.1

Source : Repec database (authors' calculations), relying on the API namsor

We observe for every article in the database the list of its authors, its year of publication, its JEL codes as well as its bibliography. We can thus recover for each registered author the list of his/her coauthors, his/her main fields of publication, the number of his/her articles as well as the number of citations received by his or her publications. The distributions of these two last variables are nonstandard, with a very large accumulation at the bottom of the distribution but also a thick distribution tail (see Figure 1 for the number of citations per author). This translates into large differences between the average number of citations, almost 7 citations per article, and its median, only 2 citations (Table 2). Similarly, the number of articles per

[3]These figures can be compared with those provided by the 2015 Statistical Report on the Status of Women in the Economics Profession of the Committee on the Status of Women in the Economic Profession, based on a survey of 124 U.S. doctoral departments and 126 non-doctoral departments. The share of women amongst all faculty members is 22% in the economics departments with doctorate, and 34% in those without oneXXX vÃľrifier doctorate. This distinction cannot be made here.

author is 12 on average with a median of 6. Women in the database have published less articles and are less cited than their male counterparts. This may partly reflect the fact that they have on average a lower professional experience.

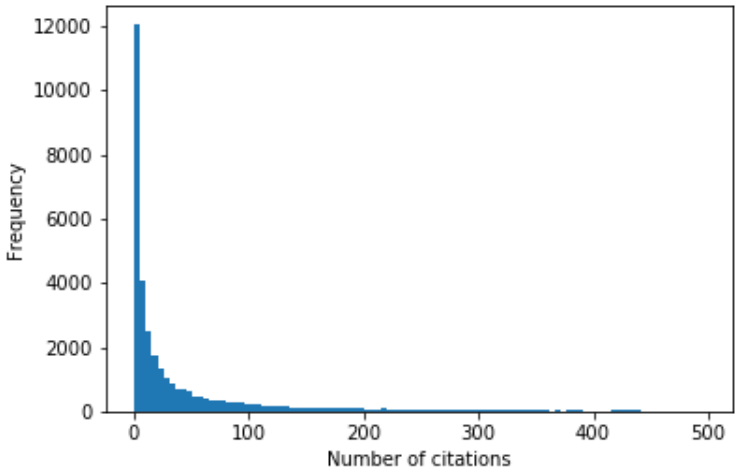


Figure 1: Distribution of citations received per author
Source: Repec data (authors' calculation)

Table 2: Researchers productivity: number of articles and average citations per article

	Number of articles per researcher			Number of citations per article		
	Mean	Median	Maximum	Mean	Median	Maximum
All	12	6	391	7.1	2.3	490.5
Gender						
Men	13	7	391	7.7	2.5	490.5
Women	8	4	197	5.5	1.7	201
Unknown	9	4	181	5.7	2.1	168
Experience (years)						
[0- 10]	4	3	116	3	1	169
[10-20]	11	8	207	8	3.4	336.4
[20-]	28	20	391	13.8	5.8	490.5

Source: Repec database (authors' calculation)

3.2 Research network description

The RePEc database allows us to analyze the links between researchers, represented as networks. The academic network can be seen as a graph where researchers form the nodes (or vertices), and their relationships are symbolized by the edges (see for instance (Jackson, 2008) for a detailed description). Specifically, an edge indicates here that the researchers have coauthored an article. Several indicators have been proposed in the literature to describe a network and identify the most influential or central nodes. A pair of researchers is considered as indirected connected in the graph if they have not coauthored but one of the pair has coauthored an article with a researcher who has coauthored with a researcher[...] who has coauthored with the other member of the pair. Using the network terminology, it is said that there is a path in the graph between these two researchers. Several paths may actually exist between two researchers. The distance between the researchers is therefore defined as the length of the shortest path between them, meaning the minimum number of intermediaries that would be required to present these two authors to each other (a notion sometimes designed as the minimum degrees of separations between them).

In 2015, 67% of registered authors in RePec are - at least - indirectly connected. This corresponds to the main "connected component" of the academic network (26,523 authors), meaning the largest subgroup where every researcher is indirectly connected to the others. The maximum distance between two random authors within this subgroup is 21, while average distance between two authors is around 6.[4] The size of this connected component has extended over time: it was less than 5% thirty years before (Figure 2). However, in 2015, 26% of the researchers have still never coauthored with another registered author.[5] The proportion of isolated nodes was 96% at the beginning of the period.

This is consistent with Card and DellaVigna (2013), who note that while three quarters of the articles published in the five most prestigious economics newspapers[6] at the beginning of the seventies were written by a single au-

[4]This can be compared to the popular concept of the six degrees of separation, stating that every person in the world is six or fewer steps away from any other. Here we consider a much stronger connection (having coauthored a research article), than a simple contact (usually considered in this theory).

[5]The remaining of the sample are small groups of a few researchers connected between them but not with any researcher belonging to the main connected component.

[6]The selection covers American Economic Review (AER), Econometrica (ECA), the Journal of Political Economy (JPE), the Quarterly Newspaper Of Economics (QJE), and the Review of Economic Studies (RES).

thor, this proportion was cut by half in the early 1990s and by four in the early years of 2010.

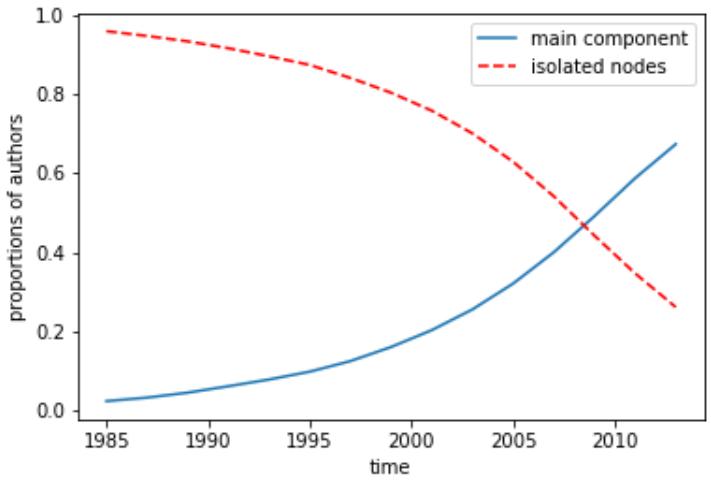


Figure 2: Share of isolated authors and main connected component
Source: RePEc database (authors' calculations)

Besides, we observe that the collaboration rate between researchers from distinct institutions has also increased over the period. This may be due to the progress in the information and communication technologies, making work easier for distant researchers. For the sake of illustration, we represent the network corresponding to French affiliations (that is small enough to be graphically represented) in 1990 and 2010. In 1990, the most connected French affiliations were PSE, AMSE and TSE. They used to cultivate partnerships with fifteen French affiliations (Figure 3). International relations were reserved for a few French affiliations and mostly with Western Europe and North America.

Twenty years later, in 2010, the trio of the most connected French affiliations remains the same, but their researchers collaborate with much more distinct affiliations (more than fifty collaborations with French affiliations and reinforced international connections). Many affiliations previously isolated (and not represented for readability issues) are now connected to the main component. In addition, collaborations with foreign countries have been democratized and diversified, with an opening towards other continents such as Asia (Figure 4).

This intensification of relations between affiliations has consequences for researchers, since it offers higher opportunities for disseminating their work within a wider community.

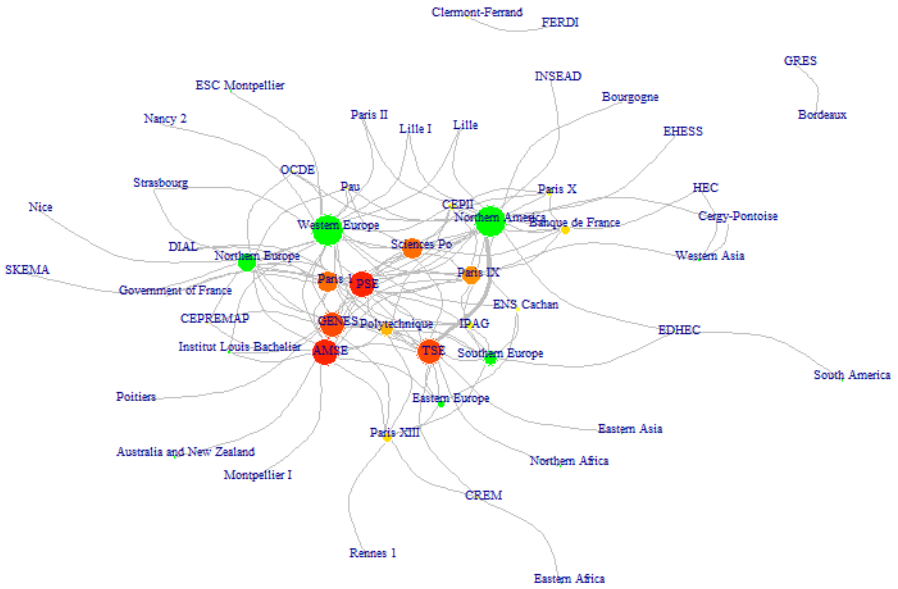


Figure 3: Network of the French affiliations in 1990

The "centrality" of a researcher in the network can be measured using several dimensions. The first measure is the number of direct neighbors of a node in the graph (usually referred to as the "degree"), corresponding here to the number of direct coauthors of a given author. We observe again that the proportion of authors with no coauthor is high, but also that some authors have more than thirty coauthors (see Figure 5).

A second common useful measure of centrality is the closeness centrality indicator. In practice, it is calculated as the inverse of the average distance from this researcher to every other researcher she/he is connected with.[7] For one researcher, it thus takes into account all the researchers who are indirectly linked to him/her and not only those she/he is directed linked with. Using this indicator, an author is considered as more central when she/he works with researchers who have many coauthors themselves (and not only if she/he has many direct coauthors, as measured by the previous indicator). Finally, a third measure is the betweenness centrality indicator. It charac-

[7]It can be interpreted as a measure of the dissemination speed of an information emanating from the node within the entire network.

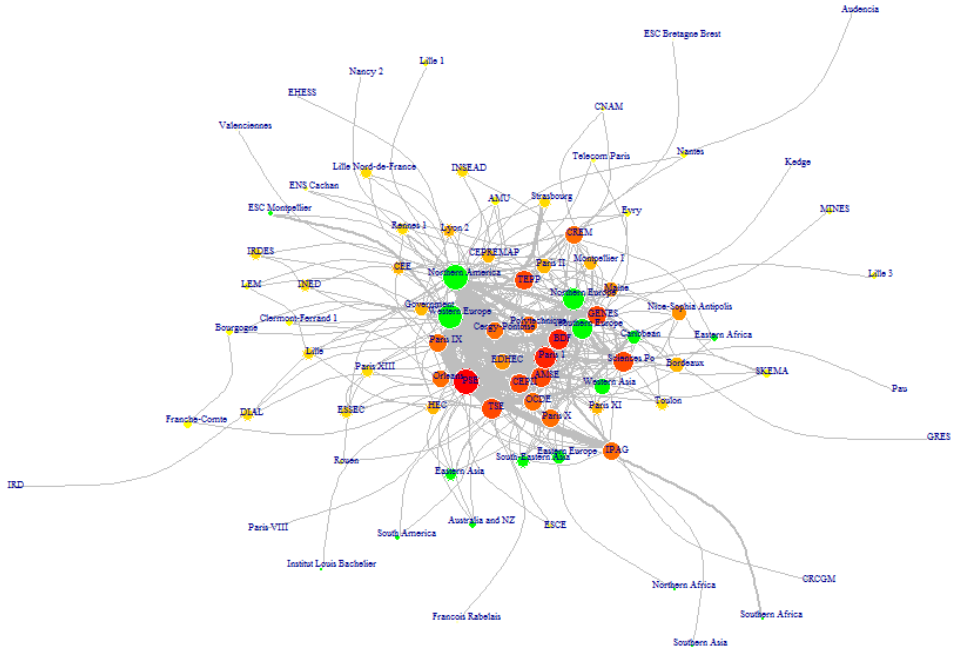


Figure 4: Network of the French affiliations in 2010
Source: RePEc database (authors' calculations)

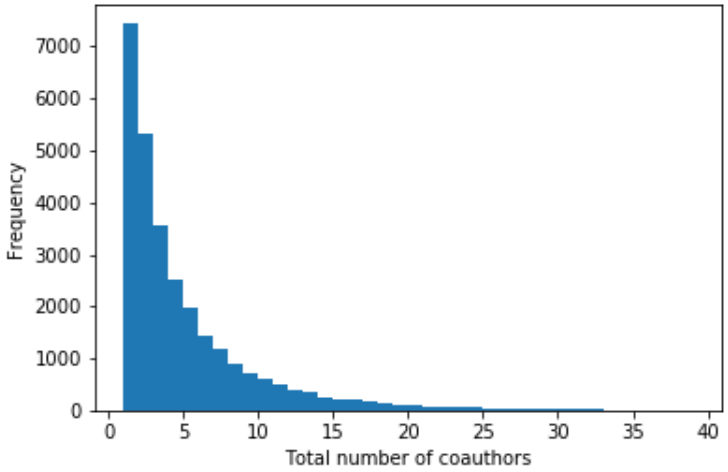


Figure 5: Distribution of the number of coauthors of each author in the sample

Source: RePEc database (authors' calculations)

terizes researchers that are at the intersection of several communities. It points out "crossing points" in a graph. In practice, for a given researcher, it is calculated as the proportion of the shortest paths between any pair of researchers in the graph that passes through this specific researcher.

For the sake of illustration, Table 3 provides a list of the most prominent authors, using different centrality indicators. Specifically, we list the top ranking authors in terms of degree, proximity and betweenness respectively. Reassuringly, the lists intersect, but the precise rankings differ depending on the chosen indicator. While these indicators are highly correlated, they indeed display complementary information.

Table 3: Top ranking authors according to different indicators of centrality

Degree	Betweenness	Closeness
Thisse, J. F.	Blundell, R.	Long, N. V.
Blundell, R.	Postlewaite, A.	Blundell, R.
Stiglitz, J. E.	Stiglitz, J. E.	Postlewaite, A.
Franses, P. H.	Tirole, J.	Dolado, J. J.
Postlewaite, A.	Woodford,M.	Thisse, J. F.
Nijkamp, P.	Blanchard, O.	Attanasio, O.
Blanchard, O.	Attanasio, O.	Baillie, R. T.
Heckman, J. J.	Chiappori, P.	Stiglitz, J. E.

Source: RePEc database (authors' calculations)

3.3 Productivity measures

As it is common in the related literature, we based our measure for the researcher productivity on the citations he/she received for his/her publications. This is usually considered as a measure of the quality of the research outputs. It may at least corresponds to the visibility of his/her works. Specifically, we propose to rely on the PageRank[8] indicator. A high PageRank means that an author is not only highly cited, but also that he/she is cited by highly-cited researchers.[9] This indicator is calculated using the directed graph defined by citations (in this case journals or authors are nodes, and

[8]Originally developed by Larry Page, it was used in particular by Google to determine the order in which web pages appear based on their relevance to a query.

[9]West et al. (2010) propose an eigenfactor index that relies on the same principle.

links are built from citations of one article by another). The PageRank is defined recursively: the PageRank of a node u depends on the PageRank of all the nodes v that point to (i.e. cite) u. Formally, it can be written as:

$$PR_u = \frac{1-d}{N} + d\sum_{v=1}^{k} \frac{PR_v}{C(v)}$$

where N is the total number of nodes of the graph, k is the number of nodes v that point to u, $C(v)$ is the number of degrees of v and PR_v is the PageRank of v. d is a damping factor which can be set between 0 and 1. In practice, the PageRank is calculated using an iterative algorithm.

Figure 6 illustrates the citation network at the level of academic journals. Here a node corresponds to one review. Its size is proportional to the total number of citations received by the aggregation of all articles of this review, and its color intensity is proportional to the PageRank. One can easily check that top ranking reviewed here are quite similar to the ones obtained by using usual bibliometric indicators. For instance, top ranking reviews as Econometrica, the Journal of Political Economy and the Review of Economic Studies (see for instance in Kalaitzidakis et al. 2003 or Chang et al. 2011), obviously stand out as ones of the most influential journals in terms of PageRanks.

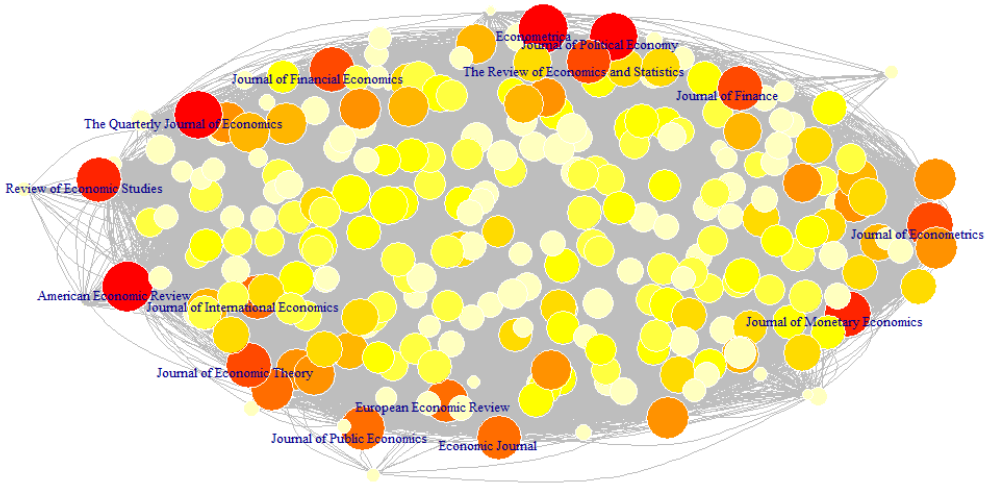
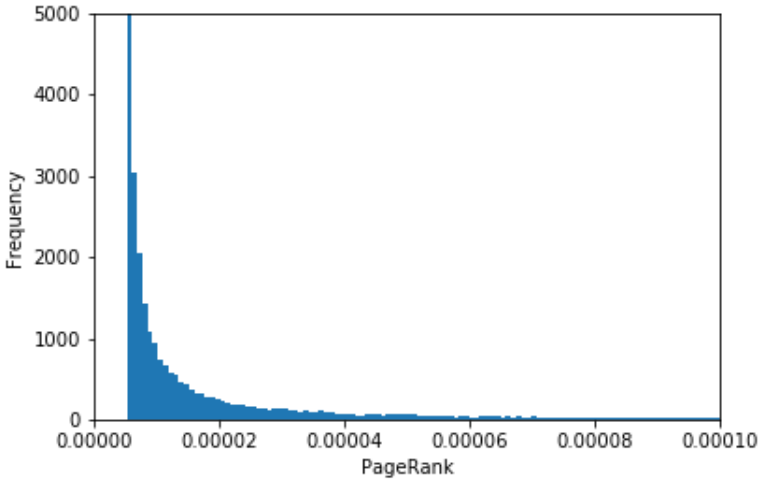


Figure 6: Most influential journals according to their PageRanks

10

The distribution of the PageRank at the author's level is highly skewed: a high proportion of authors exhibit low PageRank since a very small proportion received a very high number of citations (see Figure 7).

Figure 7: Histogram of the author PageRanks



11

4 Network formation

4.1 Econometric models

We model the probability of a common publication for two researchers i and j at year t, providing that they have not coauthored on another article before. As we are interested in the network formation, the first collaboration between two researchers may be interpreted as the creation, or at least the concrete expression, of a "strong" tie between them.

Formally, let $y_{i,j}^t$ be a dummy variable indicating that the researchers i and j have published an article together, and respectively t_i^0 and t_i^1, t_j^0 and t_j^1, the years of first and last publications of the researchers i and j, considering researchers that have been active more than just one year. We consider only pairs of researchers (i, j) whom periods of activity intersect. Our variable of interest is then:

$$P(y_{i,j}^T = 1 | X_{i,j}, y_{i,j}^t = 0 \text{ for } max(t_1^0, t_2^0) \leq t < T \leq min(t_1^1, t_2^1))$$

where $X_{i,j}$ corresponds to the characteristics of the researcher pair.

Our main covariate of interest is a covariate capturing gender-related bias in the creation of a new link. Specifically, we introduce dummies capturing whether the researcher pair is mixed-gender, male single-gender (two men)

or female single-gender (two women). We also control for the main determinants of the probability of working together, as emphasized by Fafchamps et al. (2010). We thus control for several characteristics, including measures of network proximities as well as measures of the match quality (average experience or productivity for instance). The latter is related to the benefit expected by the two researchers from a collaboration, while the former may be related to information issues (see Fafchamps et al. 2010 for a more complete discussion). Even before evaluating the mutual gain of collaboration, a prerequisite for engaging in a common work is to be mutually aware of the existence of each other.

Following Fafchamps et al. (2010), we rely on two indicators for measuring the proximity between two researchers in the coauthor network. First, the inverse of the distance between them (as defined again by the length of the shortest path between them[12]). In the following, this indicator is defined as the closeness between the two researchers. Second, the number of distinct shortest paths that connect the considered researcher pair - that may be considered as the intensity of the connections between the pair (this variable is introduced in log).

In practice, for each year t, these two measures of proximity are estimated considering the coauthor network for articles published between 1 to 10 years before the year t considered (as one may assume that the impact of previous collaborations may probably vanished over times, if not "reactivated" by new works).

The quality of the match is measured accordingly to several dimensions: the researchers' productivity (measured by the PageRank), the number of their previous coauthors (as a measure of the integration within the academic network) and their experiences. These variables are averaged over the pair of researchers, their difference in absolute level are also introduced in the model. We also use the cosine similarity between the JEL codes of their previous production, as an indicator of the overlap between the main research fields of the two researchers. As this relation is not expected to be linear, we use the variable in level and squared.

In order to control for the positive trend in the coauthorship in academic

[12]By convention, the shortest path is infinite for two authors who are in two distinct connected components of the graph. The measure of proximity between them is thus null in this case.

communities (co-authored articles are more and more numerous compared to single-authored ones, see descriptive statistics in Section 3.1), year dummies are also included.

However, the probability of working together is also related to variables that are in part unobservable. For instance, being native in the same language, having worked in the same institutions in the past, or simply sharing the same personal interests. These variables are expected to be linked with the set of observable covariates we use in our model - and notably, those corresponding to network proximity. This would result in biased estimates. In Fafchamps et al. (2010), this issue is addressed using pairwise-specific fixed effects to control for this unobservable heterogeneity in pairs of researchers. Fafchamps et al. (2010) indeed are primarily interested in quantifying the role of the network impact in new link creations. However, in doing so we could not identify the impact of our main variable of interest, the fact that the pair is single or mixed-gender, which is time-invariant.

We thus rely here on a Chamberlain-Mundlak approach, using specific pairwise random effects, year dummies and measures of the main covariates averaged over the period observed. However, this common specification has to be adjusted. Specifically, following Wooldridge (2016), we estimate a heteroskedastic Probit model using as complementary variables the pair-average of observables *interacted* with years dummies (over the whole period), and as explanatory variables in the variance year dummies. Wooldridge (2016) has shown that, providing that the link function is a Probit and with a normal distribution for the random effects, this specification coincides with the random effect one. The underlying assumption is that the unobservable characteristics that matter for the creation of a productive link are not related to the gender composition of the pair.[13]

In the main specification, we model the creation of a new link between a pair of researchers i and j. It is fully symmetric (we cannot distinguish the impact of the researcher i's characteristics - for instance, being a man or a woman- at the moment of the link creation). To go a step further, we use alternative specifications that interact some specific characteristics of one re-

[13]Another drawback of the fixed-effect specification here is that it relies on conditional logistic specification to get rid of the fixed effects (as we model a binary outcome). In practice, estimation relies only on a subset of the sample, those for whom we observe different states. In our setting that means that the estimation sample is restricted to pairs of researchers who have actually collaborated for the first time during the period. Such a sample restriction is not required when using random effect specifications.

searcher to the pairwise characteristics. In practice, we set one characteristic of the researcher i to a given value, and model the probability of a new link with the researcher j given the characteristic of this researcher and other pair-related characteristics. For instance, we restrict the sample of pairs constituted by a woman i, and evaluate the probability of coauthorship with a researcher j depending on the gender of j, once controlled for the same set of pair characteristics as before.

Because of our large sample size it is not computationally possible to estimate a model on all possible pairs of authors. We have indeed 30,324 authors active more than one year in our database, and the number of potential collaborating pairs is thus about 459 millions. Aside from computational issues (due to the sample size), as the number of "actual" collaborations between pairs is small, neglecting to address the statistical issues raised by modeling very rare events may lead to biased results. Following King and Zeng (2001), we sample all variable events (in this case, actual collaborating pairs) and only a proportion of the (very numerous) non-events (here, the non collaborating pairs among possible pairs at a given time), which provides a sample of 271,457 pairs.

4.2 Results

Our results suggest a slight gender-related bias in the network creation. A new cooperation is less likely when the two authors are of different gender that when they are of the same gender, once controlled for the main determinants of a new link creation (see column (1) in Table 4). Moreover, a new connection is much more frequent between two female researchers than between two male researchers. We observe that this bias is specific to female researchers. When interacting gender with the covariates in the model and running separate analyses for male and female researchers, we observe that a female researcher is significantly less likely to coauthor with a male than with a female researcher (see Column (2) in Table 4). This is all the more surprising that the male researchers are much more numerous amongst academic economists. For this reason, one could expect that the probability of a new link with a random researcher would be more frequent with a man than with a woman. When restricting the sample to pairs constituted by a male researcher i, we do observe that he is more likely to establish a new connection with a man than a woman, but in this case it could be simply a direct consequence of the under-representation of women amongst economists. Indeed, we observe a positive gender-related bias in favor of male

in the probability of coauthoring for male researcher i that is statistically significant only in his first years of experience (see Table 5). This may reflect the fact that many young researchers co-write their first papers with their PhD supervisors, and that male researchers are over-represented amongst this population. For young female researchers we observe a significant bias against coauthoring with male researchers, but of much smaller magnitude than for their more experienced female colleagues.

We also observe that a new connection between two researchers is all the more likely that they are productive (as measured by their average PageRank, see Table 4), but also that they do not differ too much in this dimension. These factors are more important for female than male researchers.

Concerning the other covariates, our results comfort, as observed by Fafchamps et al. (2010), that both the proximity of these researchers within the researcher network and some kind of complementaries between them make the creation of a new link between them more likely.

As an illustration of the former, we observe that the number of the shortest paths between these two researchers has a positive impact on the probability of coauthoring for the first time. This may be interpreted as the fact that the higher this number, the higher the opportunities for the pair to be connected. As suggested by Fafchamps et al. (2010), the expected impact of this variable is theoretically ambiguous. In one hand, the probability of being mutually introduced increases with the number of different paths that connect this researchers. On the other hand, being connected to a large variety of networks is obtained by multiple collaborations, some of which may still be in progress, therefore limiting the time available to launch new ones. Fafchamps et al. (2010) show that the second effect predominates in their data while the first effect predominates here. However, when analyzing separately male and female researchers, we observe that when the researcher i is a woman the probability of coauthoring with the researcher j is not correlated with the number of connections between the pair.

A new link is also more likely when researchers complete one another in some dimensions. For instance, the difference in the number of coauthors within the pair of researchers is highly positively correlated with the probability of a new link. An important difference seems to make a match appear as mutually profitable. For instance, a young researcher may be willing to collaborate with a well-connected mentor, who in return values the high time availability of his/her colleague. On the contrary, the probability of a new match declines with the number of existing links - as suggested by the nega-

Table 4: Determinant factors of the probability of coauthoring for the first time (heteroscedastic Probit with random effects)

		Women	Men
Intercept	-1.807^{***}	-1.879^{***}	-1.835^{***}
	0.029	0.101	0.030
Composition of the pair			
Single-Gender Pair (men)	-0.016^{***}		
	0.003		
Mixed-Gender Pair	-0.020^{***}		
	0.003		
Gender			
The coauthor is man		-0.014^{***}	0.010^{***}
		0.003	0.002
Proximity in the network			
Inverse distance	0.039^{***}	0.037^{***}	0.039^{***}
	0.003	0.006	0.003
Link intensity	0.007^{***}	0.002	0.007^{***}
	0.001	0.002	0.001
Characteristics of the pair			
Average number of coauthors	-0.057^{***}	-0.069^{***}	-0.057^{***}
	0.004	0.012	0.004
Absolute difference in the number of coauthors	0.199^{***}	0.200^{***}	0.202^{***}
	0.012	0.033	0.013
Average Page Rank	0.021^{***}	0.059^{***}	0.020^{***}
	0.003	0.014	0.003
Absolute difference in PageRanks	-0.030^{***}	-0.072^{***}	-0.028^{***}
	0.003	0.015	0.003
Average experience	-0.012^{***}	-0.014^{***}	-0.011^{***}
	0.001	0.003	0.001
Research field overlap	0.125^{***}	0.125^{***}	0.123^{***}
	0.008	0.021	0.008
Research field overlap (square)	-0.052^{***}	-0.055^{***}	-0.051^{***}
	0.004	0.010	0.004

Source : RePEc database. Note: Heteroscedastik Probit estimation. The pair-average observables variables are interacted with year dummies, and the scale also depends on year dummies (see Equation 4). Continuous variables are standardized. The model includes year dummies, pair-specific random effects (normal distribution), pair-average variables over the period of observations (Proximity in the network, Log Shortest Paths, absolute differences in the number of coauthors, average PageRank, Absolute difference in PageRank, research field overlap in level and square). The pair-average observables variables are interacted with year dummies, and the scale also depends on year dummies (see Equation). $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

tive estimate corresponding to the average number of previous coauthors of the pair. Because of time constraints it could be more difficult to invest in a new project.

We also observe that working on close research fields (measured by the first digit JEL code) has a positive effect, but that this impact decreases with the overlap intensity (as researchers may also look for complementary skills). Indeed the probability of working together is expected to increase with their proximity in their research interests. However, one researcher may also seek complementary skills to his/hers. Consequently, it could be less attractive to collaborate with someone who works on the very same subfields.

Table 5: Determinant factors of the probability of coauthoring for the first time by level of experience

	[-,5]		[5,10]		[10,-[	
	Women	Men	Women	Men	Women	Men
Intercept	-2.130^{***} 0.202	-1.800^{***} 0.059	-2.641^{***} 0.478	-1.869^{***} 0.069	-2.454^{***} 0.332	-1.920^{***} 0.041
Gender						
The coauthor is a man	-0.008^{**} 0.003	0.015^{***} 0.003	-0.026^{***} 0.009	-0.003 0.003	-0.026^{***} 0.009	-0.001 0.002
The coauthor is a woman	*Ref*	*Ref*	*Ref*	*Ref*	*Ref*	*Ref*
Proximity in the network						
Inverse distance	0.027^{**} 0.011	0.049^{***} 0.006	0.041^{***} 0.012	0.038^{***} 0.005	0.041^{***} 0.011	0.036^{***} 0.003
Link intensity	-0.004 0.003	0.013^{***} 0.003	-0.002 0.005	0.005^{**} 0.002	0.003 0.005	0.005^{***} 0.001
Characteristics of the pair						
Average number of coauthors	-0.091^{**} 0.037	-0.163^{***} 0.018	-0.101^{***} 0.026	-0.113^{***} 0.012	-0.072^{***} 0.018	-0.053^{***} 0.004
Absolute difference in the number of coauthors	0.239^{**} 0.096	0.454^{***} 0.049	0.317^{***} 0.076	0.340^{***} 0.035	0.215^{***} 0.048	0.202^{***} 0.014
Average Page Rank	0.106^{**} 0.045	0.130^{***} 0.018	0.099^{***} 0.037	0.057^{***} 0.011	0.051^{**} 0.021	0.016^{***} 0.003
Absolute difference in PageRanks	-0.102^{**} 0.043	-0.130^{***} 0.017	-0.089^{***} 0.034	-0.043^{***} 0.009	-0.066^{***} 0.023	-0.023^{***} 0.003
Average experience						
Research field overlap	0.115^{**} 0.046	0.242^{***} 0.026	0.104^{***} 0.029	0.099^{***} 0.012	0.066^{***} 0.020	0.084^{***} 0.007
Research field overlap (square)	-0.053^{**} 0.022	-0.126^{***} 0.014	-0.027^{*} 0.015	-0.016^{***} 0.006	-0.015 0.013	-0.022^{***} 0.003

*Source: RePEc database. Note: Heteroscedastik Probit estimation. The pair-average observables variables are interacted with year dummies, and the scale also depends on year dummies (see Equation 4). Continuous variables are standardized. The model includes year dummies, pair-specific random effects (normal distribution), pair-average variables over the period of observations (Proximity in the network, Log Shortest Paths, absolute differences in the number of coauthors, average PageRank, Absolute difference in PageRanks, research field overlap in level and squared). *p<0.1; **p<0.05; ***p<0.01*

5 How does the network impact on academic citations

The creation of a new link in the academic network may have direct effects on productivity. The knowledge production is more and more the result of a collective effort. In this section, we question another indirect outcome of collaboration. While researchers are mostly evaluated on quantitative indicators such as citations rates, we test whether a good network insertion may, by itself, increases the factor impact of one's work.

Our variable of interest is, for each researcher, the number of citations (excluding self-citation) received by an article he/she has published at year t over the subsequent five years.[14] We are mainly interested in the impact of network measures, namely the degree (number of direct coauthors), closeness and betweenness centralities (corresponding respectively to the fact of being indirectly connected to many researchers, and of being connected to several distinct academic communities). These three indicators (calculated using the network of coauthors) summarize the position of a researcher, viewed as the intensity of the links he/she establishes with his/her peers, the size of the community he/she has access to and his/her role in its coherence. Provided by the analysis of the network of coauthorships between researchers, they therefore correspond to a measure of professional activity but not directly of productivity.

Modeling the citation rate received by research articles faces the problem that the distribution is highly skewed. We observe both an accumulation around zero and the presence of very large values. These features are poorly taken into account by simple OLS. Quantile regressions are a more suitable tool in this setting. Moreover, modeling the deciles makes it possible to refine the analysis, considering the impact of different factors not only on average, but also by determining whether they contribute to widen the gaps between the most and least cited articles.

However, unobserved heterogeneity (for instance, the fact of having started his/her professional career into a prestigious affiliation) may also impact our outcomes. We again take advantage of the longitudinal dimension of our data to control for this individual heterogeneity. In practice, we rely on the

[14] When an author has published several articles in the same year we average the 5-year citation numbers over these articles.

fixed-effect method proposed by Koenker (2004). Formally, the model is:

$$Q_\tau(y_{it}|\alpha_i, X_{jt}) = \alpha_i + X_{jt}\beta_j^\tau \tag{1}$$

$$min_{\alpha,\beta} \sum_\tau \sum_i \sum_t \rho_\tau(y_{it} - \alpha_i - X_{jt}\beta_j^\tau) + \lambda \sum_i \alpha_i \tag{2}$$

where ρ_τ is the check function usually introduced for estimating quantile regression. λ is the parameter that penalizes the fact of estimating several parameters and has to be set to a value. When this value is high, the estimation is close to a regression that neglects the individual fixed-effects. With λ set to zero, the model tries to estimate all the individual effects, but it will induce high variability in the parameter estimations. In practice, we estimate several specifications and we verify that our conclusions concerning the impact of our main variables of interest are robust when using different values of λ. We also control for several time-varying variables corresponding to quality, proxy of previous productivity, institutions, etc.

Even when controlling for author specific effects and characteristics of past productivity, the work of a researcher is all the more cited that this researcher is well connected. All the three main indicators of the insertion in the network (number of degrees, betweenness centrality and closeness centrality) have a significant impact on the numbers of citations (see Table 7). However, the magnitude differs depending on the indicators, and varies greatly along the distribution of citations (see Figure 8) except for the number of degrees, meaning the number of his/her direct coauthors on his/her articles published before the one considered, that has a decreasing impact along the distributions. Having a lot of coauthors may reduce the risk for one's work to be just another paper in the crowd. Previous coauthors are more likely to be aware of the existence of a new paper and thus to cite it. They also may be a way of making this research known, for instance by proposing to present early versions in academic seminars. Having close connections to several researchers may be correlated to occasions to discuss one's work, to have his/her work known and thus increases the probability of being cited. This slight "bonus" provided by numerous coauthors tends to decrease along the distribution of citations, but is still observable even at the top of the distribution. For "high potential" papers, meaning those that receive the most of the citations, even when controlling for other individual characteristics of authors and journals, having a lot of coauthors brings higher citation rates.

We observe a distinct profile along the distribution when analyzing other centrality indicators. As expected the closeness centrality indicator has also

a positive impact on citation rates. Recall that this indicator measures not only whether a researcher has many direct coauthors, but also whether these coauthors are themselves closely connected to other researchers. The fact of being "not far" from many researchers has no significant impact at the very bottom of the distribution (conditionally to other covariates including the number of direct coauthors) but does exhibit an impact for the "medium" production. However, the benefit of one's community in terms of citation appears bounded. The estimates of the impact from the median to the ninth decile are very close. The estimated impact of the betweenness centrality indicator indeed suggests that at some point, to really make a difference one's has to diversify his or her connections. This indicator measures whether a researcher is connected to several distinct communities. Being connected to different communities (for instance, different affiliations) does not make much difference at the bottom of the citation distribution. However, the estimated coefficient increases with the decile. Being connected to different communities has an amplifier effect for important papers (those that would receive lots of citations). These results corroborate the importance of building an active network for researchers' visibility.

Concerning other covariates, we observe that the journal ranking captures most of the quality of the articles - as measured by the citation rate.[15] The journal PageRank has by far the highest impact on the citation rate of a researcher. The fact of having co-written in the past articles with a high-cited coauthor (a coauthor with a high PageRank) is also positively correlated with the citations one receives. Once controlled for other characteristics - including experience - the number of previous published articles has a rather negative impact on citation rates. This may suggest a quantity/quality trade-off for the researchers. The level of diversification of an author has also a positive impact on most of the distribution of citation rates. Specifically, we use an entropy index: $H_k = -\sum_k q_k ln q_k$ where q_k corresponds to the proportion of JEL code of class k affected to the articles written by this researcher before the year t in the k domain, considering that an article can be included in several domains. The higher this indicator, the more the researcher has written articles classified in various fields. However, when considering " blockbusters", meaning the very top of the distribution, a higher diversification has a negative impact. This suggests that these very cited articles are written by recognized experts in their field (which cannot be obtained by dispersion in various fields of research).

[15] All these indicators are computed using the coauthor network as observed in $t-1$ (thus just before the publication of the selected articles) to avoid endogenous effects.

Table 6: Impact of individual and networks characteristics on log citations of article

	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Intercept	-0.137^{***} (0.026)	0.263^{***} (0.029)	0.676^{***} (0.028)	1.229^{***} (0.026)	1.987^{***} (0.030)
Degrees (t-1)	0.030^{***} (0.004)	0.021^{***} (0.003)	0.022^{***} (0.003)	0.019^{***} (0.002)	0.013^{***} (0.004)
Closeness (t-1), log	0.003 (0.009)	0.115^{***} (0.010)	0.151^{***} (0.009)	0.144^{***} (0.009)	0.119^{***} (0.012)
Betweeness (t-1), log	-0.013 (0.012)	0.032^{**} (0.013)	0.059^{***} (0.012)	0.077^{***} (0.011)	0.078^{***} (0.014)
Intensity of links with co-authors (t-1)	0.146^{***} (0.013)	0.316^{***} (0.013)	0.333^{***} (0.015)	0.339^{***} (0.013)	0.328^{***} (0.015)
Number of co-authors	-0.000 (0.006)	0.004 (0.005)	0.007 (0.005)	-0.001 (0.005)	-0.001 (0.006)
Entropy	0.175^{***} (0.015)	0.226^{***} (0.014)	0.183^{***} (0.013)	0.077^{***} (0.012)	-0.085^{***} (0.018)
Journal PageRank (t-1), log	0.298^{***} (0.012)	0.431^{***} (0.010)	0.489^{***} (0.009)	0.499^{***} (0.010)	0.506^{***} (0.012)
Author PageRank (t-1), log	0.061^{***} (0.019)	0.117^{***} (0.014)	0.138^{***} (0.014)	0.151^{***} (0.014)	0.151^{***} (0.016)
Max PageRank of co-authors (t-1), log	0.124^{***} (0.017)	0.206^{***} (0.013)	0.217^{***} (0.013)	0.227^{***} (0.013)	0.241^{***} (0.015)
Number of published articles (t-1), log	-0.040^{***} (0.011)	-0.109^{***} (0.011)	-0.178^{***} (0.011)	-0.222^{***} (0.010)	-0.166^{***} (0.015)

Source: RePEc database. Note: Penalized quantile regressions (penalization parameter of the individual fixed effects λ set to 1). $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

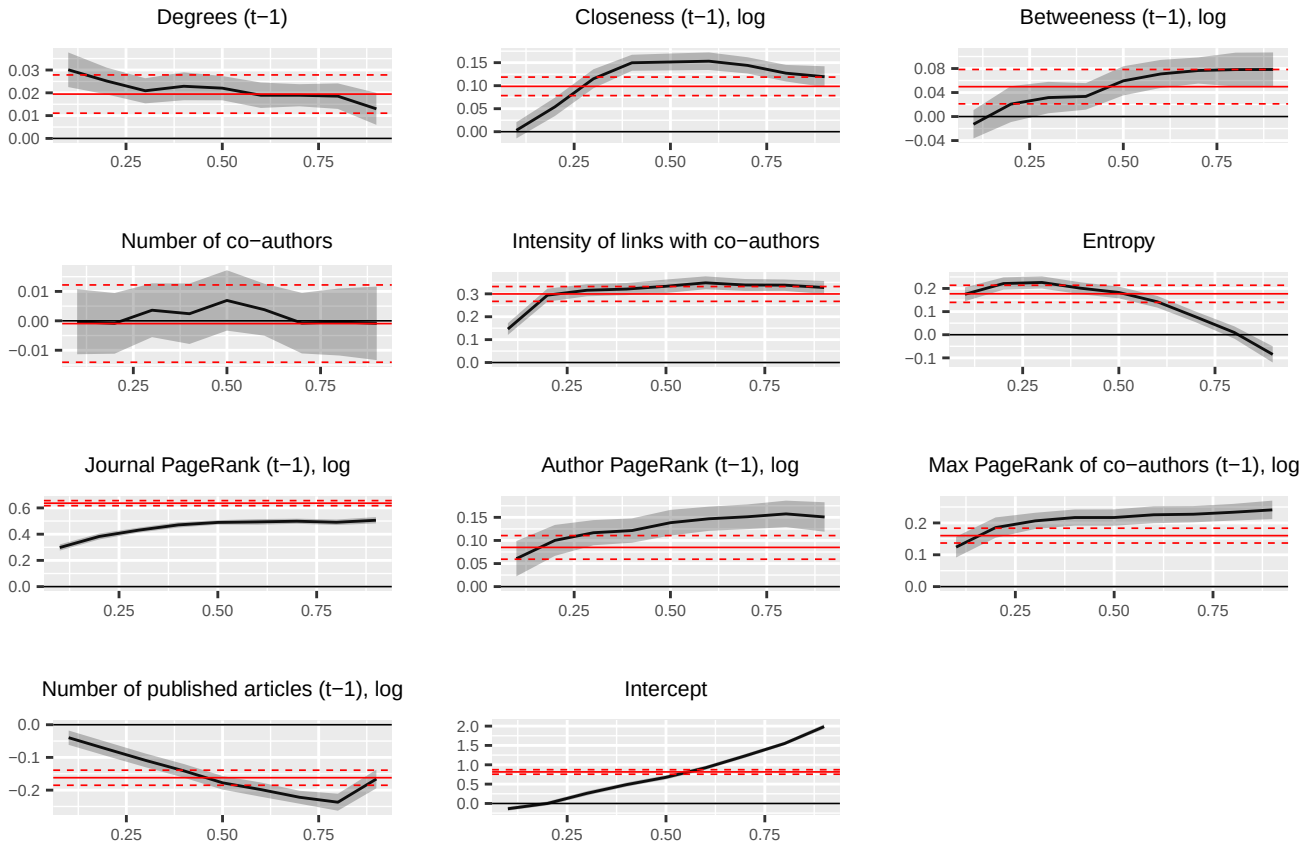


Figure 8: Coefficients per decile - estimates with fixed effects

Source: RePEc database. Estimates from a penalized quantile regressions (penalization parameter of the individual fixed effects λ set to 1). Shaded-area corresponds to confidence interval at 5xx.

When using fixed-effect parameters we face the same issue as before, as we could not estimate the impact of individual characteristics which are stable over time, and notably gender. However, we also use an alternative specification that interact every time-varying covariates with gender. According to this complementary analysis (see Table in the Appendix), we do not observe significant differences depending on gender on the return of most of variables (including those related to network). However, we observe significant positive additional positive effect of the variables related to the ranking of the Journal and those of the co-authors) on the number of citations received by female researchers compared to their male counterparts. This may corroborate the results of Sarsons (2015), who observe that women are often less personally credited for team work.

6 Conclusion

Our results suggest the existence of a gender-related bias in the creation of new link in research in economics. Based on the analysis of the articles published in the recent period, we observe that a new match between two researchers is much more likely when both researchers are of the same gender. This bias is more marked for female researchers, as they are much more likely to coauthor with another woman than with a man. This pattern may have long-term professional consequences for female researchers. As women are less numerous amongst economic researchers, one may wonder whether this homophilic behavior reduces the opportunity to integrate a research community, with cumulative impacts. As shown by our results, being already integrated in the research network may not only makes easier the creation of new connections, but may also improve the recognition and visibility of one's work. We indeed observe that the number of citations received by the articles of a researcher is positively correlated with the fact that she/he well connected in the research network.

These conclusions call for discussions. One should first emphasize that the bibliometric database used here is constituted on a voluntary-basis. The sample used for the analysis may thus not be representative of every researcher communities. This would be an issue for our main conclusion of gender-related bias matching if, for instance, female researchers who coauthor with male researchers are less prone to register on RePec than those who coauthor with female researchers. While it is not easy to find argument in favor of this assumption, the analysis of alternative sources may help to evaluate the existence and magnitude of such a potential endogenous selection bias in the RePec sample.

That said, the recent results of Sarsons (2015) provide another interpretation of the bias of female researchers against mixed-gender coauthoring. Using a database of academic economists' CVs, she measures how publications matter for obtaining tenure and observes that coauthoring, especially with male coauthors, is *detrimental* for female researchers. While male researchers do not suffer from such penalty, as coauthored and solo-authored publications of male researchers are equally considered for tenure, female researchers seem less credited than men for their contribution in team work. Sarsons (2015) also observe that such a phenomenon is less pronounced when the team is constituted only of women. This may partly explain our results, as female researchers may have less interest to coauthor with men at least in order to get tenure. However, if rational, this gender-biased preference may

also have negative side-effects. Solo-authoring, or coauthoring only with a minority reduces the opportunities to share ideas or to discuss one's work that may be helpful for improving productivity. Ductor (2015) shows indeed that, for the economists (without consideration of gender) coauthorship has a positive impact on academic productivity. Further research would be helpful to determine which effect predominates in the long term for the career of female researchers.

References

Avery, C., Athey, S., Zemsky, P., September 2000. Mentoring and Diversity. American Economic Review 90 (4), 765–786.

Azoulay, P., Zivin, J. S. G., Wang, J., 2010. Superstar Extinction. The Quarterly Journal of Economics 125 (2), 549–589.

Boschini, A., Sjögren, A., 2007. Is team formation gender neutral? evidence from coauthorship patterns. Journal of Labor Economics 25 (2), 325–365.

Bosquet, C., Combes, P.-P., 2013. Are academics who publish more also more cited? individual determinants of publication and citation records. Scientometrics 97 (3), 831–857.

Bosquet, C., Combes, P.-P., García-Peñalosa, C., 2018. Gender and promotions: Evidence from academic economists in france. The Scandinavian Journal of Economics forthcoming.

Card, D., DellaVigna, S., March 2013. Nine Facts about Top Journals in Economics. Journal of Economic Literature 51 (1), 144–61.

Chang, C., McAleer, M., Oxley, L., 04 2011. What Makes A Great Journal Great In Economics? The Singer Not The Song. Journal of Economic Surveys 25 (2), 326–361.

Combes, P.-P., Linnemer, L., Visser, M., June 2008. Publish or peer-rich? The role of skills and networks in hiring economics professors. Labour Economics 15 (3), 423–441.

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

Ductor, L., Fafchamps, M., Goyal, S., van der Leij, M. J., December 2014. Social Networks and Research Output. The Review of Economics and Statistics 96 (5), 936–948.

Dynan, K. E., Rouse, C. E., 1997. The underrepresentation of women in economics: A study of undergraduate economics students. The Journal of Economic Education 28 (4), 350–368.

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

Freeman, R. B., Huang, W., 2015. Collaborating with People Like Me: Ethnic Coauthorship within the United States. Journal of Labor Economics 33 (S1), 289–318.

Ginther, D. K., Kahn, S., 2004. Women in economics: Moving up or falling off the academic career ladder? The Journal of Economic Perspectives 18 (3), 193–214.

Hengel, E., Dec. 2017. Publishing while Female. Are women held to higher standards? Evidence from peer review. Cambridge Working Papers in Economics 1753, Faculty of Economics, University of Cambridge.

Jackson, M. O., 2008. Social and Economic Networks. Princeton University Press, Princeton, NJ, USA.

Kalaitzidakis, P., Mamuneas, T. P., Stengos, T., December 2003. Rankings of Academic Journals and Institutions in Economics. Journal of the European Economic Association 1 (6), 1346–1366.

King, G., Zeng, L., Spring 2001. Logistic regression in rare events data. Political Analysis 9, 137–163.

Koenker, R., 2004. Quantile regression for longitudinal data. Journal of Multivariate Analysis 91 (1), 74 – 89, special Issue on Semiparametric and Nonparametric Mixed Models.

Mayer, A., Puller, S., 2008. The old boy (and girl) network: Social network formation on university campuses. Journal of Public Economics 92 (1-2), 329–347.

McDowell, J. M., Singell, L. D., Stater, M., January 2006. Two to Tango? Gender Differences in the Decisions to Publish and Coauthor. Economic Inquiry 44 (1), 153–168.

Sarsons, H., May 2015. Recognition for Group Work. Working Paper 254946, Harvard University OpenScholar.

Sarsons, H., Forthcoming. Recognition for group work: Gender differences in academia. American Economic Review: Papers and Proceedings.

West, J., Bergstrom, T., Bergstrom, C., 5 2010. The eigenfactor metrics: A network approach to assessing scholarly journals. College and Research Libraries 71 (3), 236–244.

Wooldridge, J., 2016. Correlated Random Effects Models with Unbalanced Panels.

Wu, A., 2017. Gender stereotype in academia: Evidence from economics job market rumors forum. Working papers, Princeton University, Woodrow Wilson School of Public and International Affairs, Center for Health and Wellbeing.

Zimmermann, C., 2012. Academic rankings with RePEc. Working Papers 2012-023, Federal Reserve Bank of St. Louis.

Zinovyeva, N., Bagues, M., April 2015. The role of connections in academic promotions. American Economic Journal: Applied Economics 7 (2), 264–92.

A Data details: the RePEc project

The RePec project is a volunteer-driven initiative launched in the mid-1990s to create a public-access database that promotes wider dissemination of research in economics. The project maintains a database of research papers (IDEAS), articles, books and programs corresponding to about 2 million research papers from 2,300 academic journals and 4,200 collections of working papers (`http://repec.org/`) with around 45,000 authors being listed. The site is fed by publishers, research centers or directly by the authors (see Zimmermann 2012 for a more complete description). All information is freely available and structured, different types of production and versions of one paper may be easily distinguished. In order to avoid double counting, we choose to retain only published articles.

The Figure 9 illustrates the evolution of the number of authors listed in RePEc. It has strongly increased over time, but it decreased on the very last years. This may be partly due to the fact that registration is based on a voluntary basis. Since the tool has only existed for about twenty years, it is likely that it lists mostly newer researchers - and several competing social networks have emerged in the last few years. We take this time effect in the econometric analysis, by focusing on the core period.

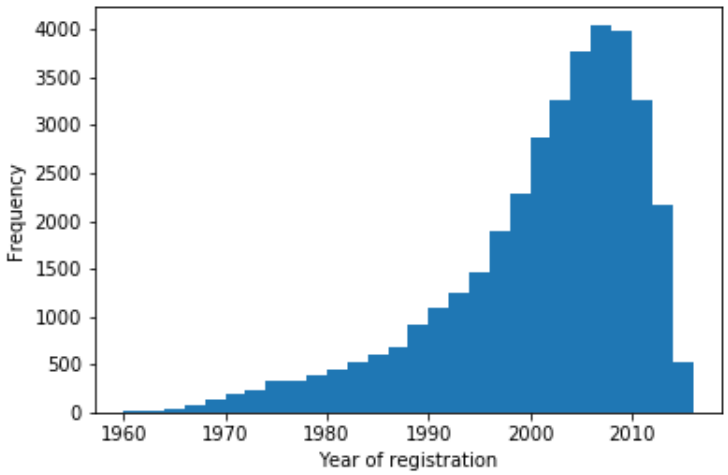


Figure 9: Newly registered authors

For each registered author, we observe the list of his or her publication (working papers and articles). We estimate a measure of his or her pro-

fessional experience using the date of the first publication recorded in the database (working paper or article). Gender is not provided by default in the database. We use the NamSor Gender API[16] to infer the gender of the authors, which is not available directly. This API infers gender from the combination of the characters present in first names and names (using classifications algorithms trained on large labeled databases). Using combination of both first name and surname improves the accuracy of the gender prediction. If the first name generally reflects the gender, the surname provides information on cultural origin. The very same first name may be mostly used for women in some culture, or men in another (for instance Andrea is a male first name in Italy, while it is often considered a female name in other countries). The surname of a person provides some information on his or her most likely origin. Accuracy is expected to be also improved by using the country of the main affiliation of the researcher.[17] In some cases, however, the sole mention of the first name and the surname is not sufficient to infer the gender of the researcher with a good accuracy. This is the case for 2% of the data.

The dataset also contains the affiliations of the registered authors. Only the current ones are known, but several affiliations may be claimed by researchers. In RePEc, researchers are supposed to weight each of her affiliations. However, no systematic rule is used for the definition of these weights. Some researchers assign the same weight to all their affiliations, others assign only one or two percent to secondary affiliations, and the highest weight to their main employer. Different affiliations usually translate into as many opportunities to present a preliminary version of a study in an in-house seminar, to disseminate it in a series of working papers, and to exchange with at least partially distinct networks researchers. Some research centers such IZA or NBER provide such a research network of affiliates - but have usually few permanent members, and we choose to consider them separately (with a specific dummies). Except from these institutions, in practice we consider for each researcher his or her most prestigious affiliation based on an indicator of centrality computed on the network of affiliations.

Every article in the database has a page that registers detailed information used for the analysis of the researcher networks. In the first place, the list of coauthors, that allows us to identify coauthorship between researchers. The article bibliography is also used for recovering citations. Using the en-

[16]This tool has been developed by Elian Carsenat, cf. http://www.namsor.com/

[17]The underlying assumption being that an author is more likely to be affiliated in her country of origin. It is not taken into account for researchers affiliated to institutions in the U.S, characterized by a high level of heterogeneity in origins.

tire database, we can link one article to all articles referenced in the database citing it.

B Supplementary estimates

Table 7: Impact of individual and networks characteristics on log citations of article

	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Intercept	-0.144^{***}	0.270^{***}	0.681^{***}	1.234^{***}	1.985^{***}
	(0.025)	(0.028)	(0.026)	(0.025)	(0.033)
Degrees (t-1)	0.028^{***}	0.021^{***}	0.021^{***}	0.018^{***}	0.013^{***}
	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)
Degrees (t-1) * Gender=F	0.025^{*}	0.009	0.003	0.005	0.002
	(0.014)	(0.015)	(0.012)	(0.013)	(0.015)
Closeness (t-1), log	0.004	0.122^{***}	0.160^{***}	0.149^{***}	0.114^{***}
	(0.008)	(0.013)	(0.011)	(0.010)	(0.012)
Closeness (t-1), log * Gender=F	−0.014	−0.031	-0.047^{*}	−0.027	0.020
	(0.023)	(0.024)	(0.025)	(0.028)	(0.033)
Betweeness (t-1), log	−0.012	0.030^{*}	0.059^{***}	0.079^{***}	0.073^{***}
	(0.012)	(0.016)	(0.015)	(0.014)	(0.016)
Betweeness (t-1), log * Gender=F	−0.015	−0.015	−0.011	−0.017	0.013
	(0.032)	(0.033)	(0.031)	(0.037)	(0.046)
Number of co-authors	0.001	0.004	0.006	0.001	0.003
	(0.006)	(0.005)	(0.005)	(0.006)	(0.008)
Number of co-authors * Gender=F	0.004	−0.001	0.004	−0.015	−0.017
	(0.015)	(0.022)	(0.018)	(0.015)	(0.015)
Intensity of links with co-authors	0.154^{***}	0.316^{***}	0.340^{***}	0.345^{***}	0.340^{***}
	(0.014)	(0.013)	(0.016)	(0.014)	(0.020)
Intensity of links with co-authors * Gender=F	−0.032	−0.020	−0.019	−0.027	−0.038
	(0.035)	(0.032)	(0.028)	(0.034)	(0.034)
Entropy	0.177^{***}	0.220^{***}	0.178^{***}	0.067^{***}	-0.104^{***}
	(0.016)	(0.015)	(0.014)	(0.014)	(0.017)
Entropy * Gender=F	0.024	0.061^{*}	0.041	0.080^{**}	0.133^{***}
	(0.031)	(0.034)	(0.035)	(0.035)	(0.049)
Journal PageRank (t-1), log	0.292^{***}	0.423^{***}	0.486^{***}	0.492^{***}	0.494^{***}
	(0.012)	(0.011)	(0.009)	(0.010)	(0.014)
Journal PageRank (t-1), log * Gender=F	0.045	0.078^{***}	0.013	0.052	0.065^{*}
	(0.038)	(0.027)	(0.026)	(0.034)	(0.037)
Author PageRank (t-1), log	0.061^{***}	0.121^{***}	0.141^{***}	0.155^{***}	0.157^{***}
	(0.021)	(0.014)	(0.012)	(0.013)	(0.015)
Author PageRank (t-1), log * Gender=F	0.157^{*}	0.108	0.205^{***}	0.187^{**}	0.171^{*}
	(0.091)	(0.084)	(0.072)	(0.082)	(0.097)
Max PageRank of co-authors (t-1), log	0.111^{***}	0.192^{***}	0.200^{***}	0.210^{***}	0.225^{***}
	(0.016)	(0.014)	(0.013)	(0.015)	(0.016)
Max PageRank of co-authors (t-1), log * Gender=F	0.095	0.122^{**}	0.172^{***}	0.138^{***}	0.158^{***}
	(0.061)	(0.050)	(0.046)	(0.049)	(0.054)
Number of published articles (t-1), log	-0.036^{***}	-0.108^{***}	-0.174^{***}	-0.219^{***}	-0.168^{***}
	(0.011)	(0.012)	(0.011)	(0.011)	(0.017)
Number of published articles (t-1), log * Gender=F	-0.055^{*}	−0.023	−0.045	−0.041	−0.022
	(0.033)	(0.035)	(0.032)	(0.033)	(0.044)

*Source: RePEc database. Note: Penalized quantile regressions (penalization parameter of the individual fixed effects λ set to 1). *p<0.1; **p<0.05; ***p<0.01*