INNOVATIVE IDEAS AND GENDER INEQUALITY*

Marlène Koffi

University of Toronto

June 23, 2021

Abstract

This paper analyzes the recognition of women's innovative ideas in comparison to the recognition received by men. Bibliometric data from research in economics are used to investigate whether there are gender biases in citation patterns. Based on machine learning techniques, I establish the similarities between papers, then build links between articles, identifying the papers citing a given paper, those cited by it, and those that should be cited. I uncover striking heterogeneity in terms of authorship. Omitted papers from references are 15 to 30 percent more likely to be female-authored than male-authored. The most likely to be omitted are papers written by women working at mid-tier institutions, publishing in non-top journals. This gender omission bias is also more prevalent when there are only males in the citing paper. Finally, I present suggestive evidence that being omitted with respect to past publications may affect an author's future productivity.

JEL codes: J16.

*Mailing address: 150 St George Street, 136; Telephone: +1 514 812 4744; Email: marlene.koffi@utoronto.ca.
I am deeply indebted to my advisor Vasia Panousi for her constant support and guidance. I am very grateful to Marti Mestieri, Dimitris Papanikolaou, Mar Reguant, Joshua Lewis, Ismael Mourifié, and Noemi Peter for extremely constructive feedback and discussions. I would like to thank Abel Brodeur, Sergio Salgado, Erin Hengel, Jonathan Guryan, Kory Kroft, Rob McMillan, the members of the Lab for Macroeconomic Policy, and the seminar participants of the University of Toronto, the Université de Montreal, Stanford IO, University California San Diego, Cornell University, the Bank of Canada, Carleton College, Max Planck Innovation & Entrepreneurship, Groningen University, the International Monetary Fund, and the World Bank, Analysis Group, Gender Online Seminars for useful comments.

1 Introduction

Women face a lower entry rate and a higher exit rate than men in industries or academic fields that require mathematical skills and analytical abilities.[1] As a result, women are underrepresented in those fields, especially in top-ranked positions (Ginther and Kahn [2004]; Ceci et al. [2014]). Preferential choices such as family, risk aversion, and competitiveness, along with discriminatory factors, have been suggested as potential explanations for this gap. Yet one mechanism has received little attention: the recognition of women's intellectual contributions.[2] Given that ideas are at the core of the research and innovation process and that being recognized and valued for one's work and ideas can be an incentive to enter or continue in a field, it is necessary to consider the question of the credit given to women's work.

This paper analyzes the state of intellectual property in academic research in economics, with an emphasis on the recognition of women's contributions. In this sense, respect for intellectual property at the academic research level arises from awareness of the individual's work and acknowledgement of relevant prior literature. Because academic research provides a suitable setting for analyzing ideas, we can test how women's ideas are perceived, used, and referred to. In that vein, this paper examines whether articles by equally relevant and innovative male and female authors are similarly credited (cited to a similar degree in the references of the articles that follow). If not, then we will postulate that a gender omission bias occurs.

[1] In 1999-2000, 13% of women received a bachelor's degree in education versus 4% for men; 2% of women received a bachelor's degree in engineering versus 12% for men (2001 Baccalaureate and Beyond Longitudinal Study, Zafar [2013]). Antecol and Cobb-Clark (2013) reach the same conclusion using survey data from the National Longitudinal Study of Adolescent Health over the period 1994-2008. Preston (1994) documents the higher exit rate of women in math-intensive fields. Hunt (2016) uses survey data from the National Survey of College Graduates to examine the difference in exit rates of women in science and engineering compared with other fields. She finds that the higher exit rate of women could be explained, to a greater extent, by women's dissatisfaction with wages and promotions.

[2] Historical examples indicate that in the past many women did not get proper recognition for their scientific achievements (Rossiter [1993]). For example, the connection between the amount of carbon dioxide in our atmosphere and climate change was made by Eunice Newton Foote in 1856. However the credit was given to John Tyndall 3 years later (see the article in the Quartz https://qz.com/1277175/eunice-foote-proved-the-greenhouse-gas-effect-but-never-got-the-credit-because-of-sexism/). A rather similar situation is described in this Forbes article in the case of Rosalind Franklin and her contributions to the DNA model https://www.forbes.com/sites/kionasmith/2018/04/16/rosalind-franklin-died-60-years-ago-today-without-the-nobel-prize-she-deserved/?sh=4f3df80179e7. Are such omissions a thing of the past, or is there a gender omission bias in modern times? Knowing whether gender omission bias is present is important for understanding and tackling women's underrepresentation.

I focus on economics within academia for two main reasons. First, the representation gap (the gap between the share of women in a given field and the share of women in the overall population) is among the largest in economics (Bayer and Rouse [2016]). Second, many voices have recently been raised against gender discrimination in economics research, which seems to be more prevalent than in other life sciences disciplines or engineering (Ginther and Kahn [2004]; Wu [2018]; Sarsons [2017]). This paper contributes to the literature by identifying a gender omission bias. It also brings a methodological contribution by constructing an omission index that identifies papers that should be cited in the references and are not. Further, it reveals heterogeneity in the omission bias, investigates some potential mechanisms, and discusses policy implications.

To achieve the research objectives, I use bibliometric data on articles published in major economic journals. The data come from the Web of Science, Econlit, and Ideas RePEc. Furthermore, textual analysis, based on big data and machine learning tools, adds valuable insights.

First, I infer the gender of the author through gender name dictionaries, and manual checking. I then extend to the "gender" of the article using the gender of the authors writing it.

Second, I build a distance measure that compares different articles and establishes a link between the citing article, the cited articles, and the articles that "should be cited" (according to the distance metric). At this level, it is difficult to establish such a relationship, especially when we know that in research, each article aims to differentiate itself from its predecessors (by style of writing, methodology, approach, etc.). One approach, followed by Zhu et al. (2015), is to ask authors which articles they think are the most important for their study. However, this approach has two drawbacks. First, survey responses may not be completely objective, but rather reflect latent biases; and second, they do not provide an obvious way to construct counterfactuals describing which authors (and thus types of authors) should be cited and are not. Besides, while it is clearly impossible to cite all the papers in the prior literature, issues arise when there is a discrepancy between authors cited and those who are not, based on their (observable) characteristics despite the relative closeness of their ideas. So we need some

measure of relative closeness to condition on. I propose an objective method that allows me to relate different articles using textual similarities. Using natural language processing, the comparison is based on textual distance tools, that link two papers based on their topical contents or the words they used. I further extend to tools based on deep learning and neural networks for textual analysis, in which word synonymy is more extensive.

Two key indices are then constructed from this analysis. The first is an omission index. I measure the propensity with which an article that is part of the relevant literature for certain articles is omitted from those articles' references. In other words, the omission index captures the fact that an article that has several similarities to a later paper is not cited in the latter. It is a dummy variable that links two papers and takes one if the citing paper has omitted the other paper from its references provided that the other paper is similar to the citing one. Consider, for example, a database with three articles: the first two articles discuss the wage gap between men and women, and the third deals with twin crises. Regarding subject matter, the relative similarity between the first two articles will be more important than their similarity with the third one. If the first article does not cite the second given their relative proximity, the omission index between the first article and the second article will be one. Otherwise, the omission index will be zero. The second index constructed is the innovativeness index, which offers an alternative way to assess the quality of an article. Similar to Kelly et al. (2018), an article is considered very innovative (and therefore of high quality) if it is new and influences future research. Given an article, the innovativeness index is a continuous variable built based on the degree of similarity with past papers in the database (papers published before this article) and future papers (papers published after this article). For instance, an article like that of Kaminsky and Reinhart (1999) about the twin crises is likely to be at the top of the innovativeness distribution because it differs from the articles before it and influences the literature after it. Unlike citation, both metrics (omission index and innovativeness index) are less likely to be biased by the willingness of some authors to cite others. Instead, they rely exclusively on the topical content. They allows me to contrast "what should be" and "what is observed" from the citation pattern.

Whereas papers in the unsupervised learning literature and, more recently, those using texts as data in economics (see Gentzkow et al. [2019] for a review) contribute to the method's external validity by presenting the relevance of textual analysis for economics, I carry out multiple internal validations to show that the methodology proposed in this article captures true patterns in the data. Further, each article's observable characteristics are combined to build an author-level database to assess the effects of the omission for the authors in terms of future publications, and to some extent, their future career trajectory.

Turning to the findings, this paper presents evidence for a gender omission bias in economics: Omitted articles from references are 15% to 30% more likely to be female-authored than male-authored. Mixed-team papers (with both male and female authors) tend to fall in between the values for both genders, ie the likelihood of omission tends to be lower than female-authored papers but a bit higher than male-authored papers. The papers most likely of being omitted are written by women working at mid-tier institutions and publishing in non-top journals. In a group of related papers, the probability of omission of those papers increases by 6 percentage points compared with men with similar attributes when the citing authors are only males. Overall, for similar papers, having female authors reduces the probability of omitting other women's papers by up to 10 percentage points. Moreover, the omission bias is much higher in theoretical fields that involve mathematical economics than in applied fields such as education and health economics. In addition, even papers written by women published in top journals are not exempted from omission bias compared to similar papers written by men.

The baseline estimation includes only articles published in top-ranked economic journals. Provided that the journal in which a paper is published is a signal of quality, this ensures that the estimation does not capture doubt regarding the articles' quality. This indicates that we are more likely to take a lower bound by controlling for the quality of an article using the quality of the journal: If there is a bias in the standards imposed on men and women (Card et al. [2020], Hengel and Moon [2020]), then the articles written by women in top-ranked economic journals will be of better quality than those written by men, and yet the bias still exists. Further, the regressions include observable characteristics for both the citing and the

cited or omitted articles, such as the affiliation of the most prolific authors, the main field, the year of publication, and the gender structure of the articles in the most similar set, the number of references. Several robustness checks (including but not limited to controlling for the methodological style, the position of the authors' names, and the extension to more than 100 journals in economics) suggest that the estimates are not overly sensitive to the choice of control variables.

To ensure that I am capturing true patterns in the data, I validate the omission index in several steps. First, the index is related to the distance metric. On average, the closer two articles are in terms of the measure of similarity, the lower the probability of being omitted (and the greater the probability of being cited). Then I run falsification tests in which I randomly generate the sample of similar papers. None of the t-statistics exceed 1.2. In other words, if the similarities suggested by the algorithm were as good as random (no true links between papers), we would not expect any bias to occur. Third, I assess whether women use more plainer language than men do, which would lead to obtaining women-men, men-women, and women-women type similarities more frequently than men-men similarities. However, empirical evaluation of this hypothesis refutes it. Finally, I use textual similarity analysis methods using advanced methods in natural language processing and neural networks, such as Glove (Pennington et al. [2014]), and Word2vec-Doc2vec (Mikolov et al. [2013]; Le and Mikolov [2014]), but also a method solely based on cross-referencing (i.e., no textual analysis). Qualitative results remain invariant regardless of the method, which confirms that this relationship is unlikely to be spurious.

I then examine several alternative explanations for the empirical patterns. For example, checking for the degree of seniority does not change the qualitative results. In particular, I do not find evidence that publication history is the central reason for the gender omission bias. Another mechanism that could explain the omission would be the lack of information; in other words, those omitted papers may not be known. To test this hypothesis, I construct a proximity index that captures the distance between two authors—if they are from the same institution, if they have already been coauthors, or have a common coauthor (peer effect). None of these measures is accompanied by a relative reduction in female-authored papers' omission compared

with male-authored papers. Further, I investigate whether the effect is specifically related to being perceived as a woman, which would present evidence of potential discrimination. To do this, I compare the omission bias between articles written by women whose names have a known feminine connotation and those whose names have a more ambiguous (androgynous) feminine connotation. I find that the former type is more likely to be omitted than the latter. The discussion that follows using a simple citation model shows that it is difficult to distinguish the type of discrimination. Moreover, this highlights the fact that higher citations in fields with more women can either be because women omit fewer other women or because men give more credit to women or "are constrained" to give them credit because women are more likely to review their articles in those fields.

Next, I present how the omission bias creates a discrepancy between the quality measured by citations count and the innovativeness index. I also provide evidence of the gender omission bias in other fields apart from economics such as mathematics and sociology. Finally, being omitted with respect to past publications reduces the probability of getting published in a top-five journal in the future by up to 5%.

2 Related Literature

Through the subjects discussed and the techniques used, this study builds on several areas of the economic literature. First, the question of whether women get enough credit, and therefore recognition, for their research is at the core of this paper. In this sense, the paper is complementary to Sarsons (2017) and Sarsons et al. (2021). Indeed, Sarsons (2017) provides key insights on the question by testing whether the uncertainty about the individual contributions of co-authors favors men in terms of tenure rates compared to women. She finds that men received more credit from coauthoring with women. I explicitly use article references to assess to whom credit is most often attributed and whether this is done to the detriment of women. Moreover, Sarsons's findings suggest that women are worse off when they collaborate with men. I show that women also fare worse when they do not collaborate with men: Mixed-

gender teams received treatment midway between that received by single-gender teams.

This paper is linked to the general literature on gender discrimination in academic research. More specifically, three main points emerge from the literature. The first is the presence of stereotypes. Wu (2018) highlighted that female authors are most often associated with physical characteristics, whereas male authors are most associated with intellectual characteristics. The second element is the difference in standards for and evaluations of men and women. For example, Hengel (2020) shows that women experience longer delays in the review process and are asked to make many more revisions before getting published. Along the same lines, Card et al. (2020) show that to publish in the same journal as males, females are required a citation premium. Similarly to Wu (2018), this paper uses textual analysis techniques to extract relevant information. I additionally construct two indices that reveal hidden patterns not identified by traditional numerical data. Further, the paper adds to the literature by arguing that in addition to higher standards and stereotypes, women also face a lack of recognition of their work even when they publish high-quality papers compared with their male colleagues. Third, this paper also addresses a point raised by Hamermesh (2018): that credit may not always go to the right person.[3]

The paper is also related to the literature on the existence bias in citation. Both the number of citations and the journal of publication are commonly used measures to evaluate the quality of a paper (Hilmer et al. [2015]; Heckman and Moktan [2020]). However, Wilhite and Fong (2012) and Fong and Wilhite (2017) show how citations may not necessarily reflect the merit of the cited article or are manipulated to increase the journal's impact factor. Citations could therefore reflect a strategic decision (Lampe [2012]) or characterize a network (D'Ippoliti [2021]).[4] At this level, this paper both departs from and complements previous literature contrasting realized citations and relevant but omitted citations, and using an alternative measure of the scientific quality of a paper. Focusing on gender, Ferber (1986); Ferber (1988); Dion et al. (2018); and

[3]For more literature on gender and academia, see Blau and DeVaro (2007); Moss-Racusin et al. (2012); Chari and Goldsmith-Pinkham (2017); Teele and Thelen (2017); Antecol et al. (2018); Auriol et al. (2019); Lundberg and Stearns (2019); Ductor et al. (2021); Hospido and Sanz (2021).

[4]Additional evidence of the networking effect can be found in Colussi (2018), who demonstrates that publications in a journal are influenced by social connections, faculty colleagues, and Ph.D. students.

Koffi (2021) show that women's papers are mostly cited by women's papers. I find the same in this paper: From an omission perspective, women's papers are more likely to be overlooked in papers by men. However, while those papers present key evidence for gendered citations, they do not provide a counterfactual of what should have been done or if there are reasons for women to be cited and there are not.

Another central question is why we care about citations or missing citations. Jensen et al. (2009); Hamermesh and Pfann (2012); and Ellison (2013) argue that citations are important in determining labor market outcomes: They signal reputation and are crucial with respect to hiring, salaries, tenure, and grants. In Gibson et al. (2017), citations matter, especially in the low-ranked department. In line with those findings, this paper further shows that being omitted influences an author's future publication possibilities, because such authors tend to have a lower chance of publishing in top economic journals.

Last, like Kelly et al. (2018); Hofstra et al. (2020); and Koffi and Panousi (2021), I rely on document topical content to build an innovation-related index. Also, whereas Kelly et al. (2018) and Koffi and Panousi (2021) focus on patents and Hofstra et al. (2020) on scientific dissertations, I look at publications in economics journals. Furthermore, I construct an omission index, which is a distinct methodological contribution of this paper.

The remainder of the paper is organized as follows. Section 3 describes the data. Section 4 presents the procedure to construct the similarity and the omission index. This section ends with some descriptives on omission and gender. Section 5 presents the main empirical strategy and the results. Section 6 discusses some potential mechanisms and shows additional results. Section 7 assesses the effect of omissions on future productivity. The paper concludes with a discussion of implications of the empirical results.

3 Data description

The raw data are collected from three main websites, the Web of Science (WoS) database, Econlit, and IdeasRepec (IR). Together, these sites are among the largest depository of academic

research in economics. This information is then organized into a novel database.

First, a corpus is created from all papers published in the top 16 journals in economics over the period 1991-2019. Details on journal ranking can be found in Laband and Piette (1994); Kalaitzidakis et al. (2003); Kodrzycki and Yu (2006); Engemann et al. (2009); Kalaitzidakis et al. (2011); Bornmann et al. (2018); Thomson and Reuters Clarivate Analytics; and IR. The full list of journals is provided in Table *I*. As is well known, published papers are submitted to a range of controls by reviewers to ensure that they contain all relevant information concerning the prior literature. In all, the sample includes the five general-interest journals traditionally considered to be the top-five —*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economics Studies*—as well as 11 prestigious special-interest or field journals. The corpus contains full-length articles and excludes proceedings papers, comments, articles of fewer than three pages, book reviews, bibliographical items, articles without references or abstracts, editorial material, letters, and corrections. Second, a set of automated web-crawling algorithms were designed to collect the following information from WoS and IR for each paper in the corpus: title, abstract, keywords, JEL codes, references, journal of publication, date of publication, authors' names, authors' institutions, and number of citations.[5] References is an important variables for the analysis, so more detail is given below about its collection. Because about 40% of authors' names on WoS consist of initials and last names, the authors' full names were validated by the *Cited Reference API* or merged with Econlit and IR.[6] Overall, the merged database contains 24,033 papers and their associated information.

Figure I shows that the number of publications with available abstracts in the sample

[5] "Citations" refers to the number of times a paper is cited by subsequent papers. Citations can also be collected from Google Scholar (GS). However, for the same paper, the number of citations in WoS and GS often differs. This is because WoS considers citations that occur post-publication and only from other published papers, whereas GS counts citations already for the working version of a paper and from any other papers, documents, or articles. In what follows, consistent with the fact that the year of publication is used as the reference year, the preferred specifications will use WoS as the main source of citations.

[6] The "complete" names obtained by Rcrossref are also sometimes incomplete (initial of the first name + last name) or correspond to the names of the first authors in alphabetical order. For example, for a paper written by Abhijit Banerjee and Esther Duflo, Rcrossref could give "Banerjee A. and Duflo, E." or "Banerjee A." as authors' names. The WoS database is merged with Econlit and IR database using the title of the article, the journal of publication, and the authors' last names.

increased fivefold between 1991 and 2018, from about 200 in 1991 to over 1,000 in 2018. This could reflect data availability and also entering of new journals over time. The number of authors has increased over time as well, from 1.7 authors per paper on average in the 1990s to about 2.2 authors per paper in the 2010s. The average number of authors over the entire sample period is approximately 2.[7]

The terminology "references" refers to the bibliographical references to related literature in each published paper. One difficulty here is that the references provided in WoS are often incompletely listed. For example, for each reference, you have the name of the first author, the year, the journal, the volume, the number of the starting page, and the digital object identifier (DOI) when available. Therefore, I linked the papers in the references to papers in the database using the DOI. When the DOI is missing, I use the other information. To control for the fact that some articles could be cited while unpublished, an additional web-crawling algorithm was designed that allows for recovery of the title of the unpublished paper at the moment of citation. A fuzzy matching algorithm with Levenshtein distance is used here to match papers. On net, the resulting database contains 914,371 references with an average of 38 references per article. This is in total and not restricted to the top 16; some of the references are to papers published in journals lower than the top 16, and are not considered. In the end, about 30% of the references (up to 50% in recent years) are to top papers in the top 16 journals, and these are the one used for comparisons.[8]

Third, the authors' institutional affiliations are classified into three categories, based on internationally acknowledged rankings of economics departments and organizations: top tier; middle tier; and lower tier. For multiple coauthors, the paper's affiliation is taken to be the affiliation of the coauthor at the highest-ranked institution. Details of the ranking procedure are provided in the appendix. Papers written by authors from top tier institutions account for 41% of the sample, those written by authors from middle tier institutions account for 33% of the sample, and authors from lower tier institutions account for 26% of the sample.

[7]Those findings are in line with Card and DellaVigna [2013].

[8]In the robustness, the database is extended to around 100 journals.

Fourth, the gender of the authors is determined via a combination of automated algorithms and hand-collection efforts. First, I employ a gender name dictionary for gender attribution, namely the *Genderize.io* API. This API is based on an large databases of names collected from the US Census, international dictionaries, and social media, calculate the probability that a certain first or last name is associated with the male or female gender. When needed, specialized algorithms enable for web searches related to particular types of names. For example, in cases in which the authors' native country can be determined, an algorithm was designed to find country-specific name-gender probabilities. For instance, in the general sample, the name "nicola" is identified as female with probability 70%, whereas in Italy it is identified as male with probability 99%. Finally, I proceed to hand collection and first-party verification. For example, for the 200 most frequently cited authors identified as female and for the 200 most frequently cited authors identified as male, the genders were verified by visual inspection of authors' personal websites and from other publicly available sources.[9] When the gender associated with a particular name with probability higher than 80%, the author's gender is considered to be determined. For names that were assigned a probability lower than 80% by the relevant algorithms, a manual search was conducted to recover the author's gender, based on personal websites, web sources that refer to the authors using their pronouns, and articles that include authors' photographs. On net, 85% of the total 22,053 authors' names were identified as male or female. The rest were related to unisex first names, such as "taylor," or to first names that consistently appear initialized (for which I conduct the manual search), or to names for which no information could be retrieved about the author. Then, almost 77% of the authors are identified as men and 17% as women.

Fifth, the "gender composition" of each team of coauthors is identified. Four categories of gender composition are defined. A paper is identified as male-authored if all the coauthors are men. A paper is identified as female-authored if all the coauthors are women. A paper is

[9]This library of names was then augmented by merging a database of inventors' names from the World Intellectual Property Organization (WIPO), which added 8 million names, and by merging the IR list of names of the top 10% of female economists (IR, January 2019, https://ideas.repec.org/top/top.women.html). I also construct my own-built gender recognition algorithm based on first names for double-checking and correcting mistakes.

identified as mixed-authored if at least one coauthor is a woman and at least one coauthor is a man. A paper is identified as undetermined if I cannot uncover the gender of at least one of the author. A paper is identified as mostly male-authored if most of the coauthors are men. A paper is identified as mostly female-authored if most of the coauthors are women.

As shown in Table II, almost 75% of the articles in the database are male-authored, about 5% are female-authored, and the remaining are authored by mixed-gender teams. Articles with at least one female author (sole- or team-authored) make up only about 20% of the database. Figure II presents the evolution of the share of papers with at least one female author. In the early 1990s, papers with at least one female author constituted only 10% of published papers, whereas in recent years this number is closer to 30%. Similarly, the share of mostly female-authored papers has increased over time. However, the fraction of published economics papers with only female coauthors (solo or all-female teams) has remained stable over time at 5%.

The main field of a paper is found following a methodology similar to Azoulay t al. 2017. I construct a classification algorithm using papers for which the first one-digit JEL is consistent over all the JEL listed. More details on this procedure can be found in the online appendix. The gender composition of the authors differs systematically across fields. For example, in labor economics and the economics of education, about 8% of the papers are female-authored and about 23% have at least one female coauthor, compared with 3% and 15%, respectively, in the fields of theory, finance, and macroeconomics. These findings are similar to those in Card et al. (2020).

Figure III presents the gender distribution of references by field of study. On average across fields, about 80% of citations are for prior papers that are "male" (the blue bars), where male indicates an all-male team (solo or all-male authors). Most of the remaining references are to prior published papers written by mixed-gender teams (gray bars). Only a tiny fraction of the references are for papers written by female teams (solo or all female, orange bars). This fraction is marginally higher in the fields of labor, education, and Industrial Organization.

Figure IV presents the gender distribution of references as a function of the gender structure of the paper's author team. The vertical axis is the gender of the citing team. The horizontal

axis is the probability of a related paper's being cited when the authors are male (blue bars, solo or all male team); female (orange bars, solo or all female team): or mixed-gender (grey bars). Comparing the orange parts of the bottom and middle bars, we can see that women are three times more likely to be cited by women authors than by men authors.

4 Pairwise similarity and omission index

In this section, I construct the pairwise similarity and the omission index using the techniques of textual analysis, natural language processing, and unsupervised machine learning. The pairwise similarity determines how similar two papers are in terms of context. This is basically a linguistic distance metric that evaluates how closely related two papers are, based on their subject matter. The "omission" index, captures the papers omitted from the references of a given paper when they should have been cited, given their relative closeness provided by the similarity.

4.1 Natural language processing

Each paper is linked to the chronological sets of preexisting and subsequent papers using commonalities in the topical content of each pair of papers. In turn, the topical content is culled from titles, abstracts, and keywords.[10] These so-called textual data are cleaned and taxonomized into sets of words. For example, cleaning involves dropping words that appear very frequently across papers, and therefore cannot be used to determine the degree of similarity across papers, such as "the" or "and."[11] Taxonomy encompasses, among other things, the attribution of parts of speech to each word that is not dropped during the cleaning stage.

The set of words includes individual words as well collocations or n-grams. Collocations are basically combinations of multiple words and occur frequently in many fields of economics. For example, "public debt" is a bigram and "capital income taxation" is a trigram. These

[10]In the online appendix, I show that taking either the abstract or the introduction or the full text of the paper do not change the relative similarity.

[11]The appendix provides details on the data-cleaning process.

collocations are identified via unsupervised machine learning, which means that they are allowed to automatically evolve over time, following the evolution of the language used by economists. For example, the collocation "idiosyncratic income risk" became more frequent after the seminal 1994 paper by Aiyagari, while the collocation "unconventional monetary policy" started being used after the onset of the 2007 financial crisis in the US.

4.2 Term frequency-Inverse document frequency

Term frequency-inverse document frequency (TFIDF) is a metric often used in machine learning to identify the relative frequency of a word in a corpus or collection of documents. The term frequency (TF) component gives the frequency of each word in the set of words in each document. Specifically, the TF is the ratio of the number of times the word appears in a document to the total number of words in that document. Clearly, the TF increases as the number of occurrences of the word within the document increases. For each word w in each paper p, the TF is therefore computed as:

$$TF(w,p) = \frac{Card(w \in p)}{Card(p)} \tag{1}$$

where $Card(w \in p)$ is the number of times the word w appears in paper p, and $Card(p)$ is the cardinal of p or the number of words in paper p.

The inverse document frequency (IDF) is the logarithm of the inverse ratio of the number of documents in which a word appears over the total number of documents in the corpus. Let C be the corpus or set of all documents in the database and $Card(C)$ the cardinal or number of papers in the corpus C. The IDF is then computed as:

$$IDF(w) = -\log\left(\frac{\sum_P \mathbb{1}_{w \in P}}{Card(C)}\right) \tag{2}$$

Thus, words that appear in every document will have an IDF equal to zero, whereas the words that occur less frequently in the corpus will have a high IDF, because they are more informative

for assessing similarities across documents. For example, for the similarity comparison of two papers, words like "taxation" or "bayesian" will be more useful than words like "paper" or "analysis", and hence they should enter more prominently into the similarity calculation.

The TFIDF of a word is then the product of the TF of the word times the IDF of the word, or $TFIDF = TF \cdot IDF$. A low TFIDF may indicate that the word appears infrequently in the document (low TF) or that it is a very common word that appears in many documents (low IDF). A high TFIDF indicates that a word appears relatively frequently in one document, but it does not appear in most other documents of the corpus; hence it is crucial for the content of this particular document.

However, the traditional IDF does not take into account the evolution of the natural language or the introduction of new vocabulary or terminology over time. Following Kelly et al. (2018), the IDF is therefore adjusted to give a higher weight to newly introduced concepts. Let $\tilde{C}(t)$ be the corpus of papers before a given date t. Then, the adjusted IDF (AIDF) is given by:

$$AIDF(w,t) = -\log\left(\frac{\sum_{p \text{ prior to } t} \mathbb{1}_{w \in p}}{Card(\tilde{C}(t))}\right) \tag{3}$$

Thus, the AIDF is the logarithm of the inverse ratio of the number of papers published before paper p in which a word appears over the total number of papers published before paper p. Basically, it is a retrospective version of the IDF. Because the AIDF varies with time and across words, it attributes importance or weight to each word depending on the degree of utilisation of the word over time. As a result, it reflects the state of the art or the frontier of innovation up to the arrival of each new paper. Clearly, the $ATFIDF = TF \cdot AIDF$.

4.3 Pairwise similarity

The pairwise similarity, which measures the textual or conceptual similarities across two papers, is basically a cosine similarity distance measure. The cosine similarity is a measure of similarity between two nonzero vectors of an inner product space. It measures the cosine of the angle between the vectors, where the cosine of a 0-degree angle is 1 and the cosine of a 90-degree

angle is 0. Each of the two papers to be compared is represented by a vector based on the ATFIDF of each word. Let U and V be the respective vector representations of papers p and p':

$$U = [atfaidf(w_1, p, t), atfaidf(w_2, p, t), ..., atfaidf(w_n, p, t)]^T$$
$$V = [atfaidf(w_1, p', t), atfaidf(w_2, p', t), ..., atfaidf(w_n, p', t)]^T$$

Then I define the cosine similarity used in this paper as, $\lambda_{p,p'}$, the relative angle between these two vectors:

$$\lambda_{p,p'} = \frac{cos(p, p')}{\max_{p' \in C, p' \neq p} cos(p, p')} = \frac{\frac{U.V}{\|U\|\|V\|}}{\frac{U.V_{max}}{\|U\|\|V_{max}\|}} \tag{4}$$

Where V_{max} is the vector representation of the paper satisfying $\max_{p' \in C, p' \neq p} cos(p, p')$ for a given paper p. Clearly, $cos(p, p') \in [0, 1]$. Papers that are similar tend to use the same words with the same frequency, so their vector representations have a trigonometric angle close to 0, and therefore the cosine similarity measure will take a value close to 1. At the opposite end, papers that have no common concepts will yield a cosine of around 0.[12]

4.4 Omission index

Next, the most similar set for paper p or the relevant prior literature for paper p, denoted by $\mathcal{P}_p$, is defined as the n-papers, denoted by p_i, with the highest cosine:

$$\mathcal{P}_p = \{p_1, p_2, ..., p_n\} \text{ such that for } i \in [0, n], \lambda_{p,p_i} > \lambda_{p,p'}, \quad \forall p' \in C \backslash \{p_1, p_2, ..., p_n\} \tag{5}$$

The preferred specification uses $n = 5$, and thereby examines which of the top five most-related prior papers are omitted from the references of the paper under consideration. However, the qualitative results are robust to higher values of n. Next, the omission index for the comparison between similar papers p and p', of which p' was published first, is a binary variable,

[12]In the analysis that follows, similarities below 0.05 are set at zero, and zeros are dropped, to reduce the computational burden and remove extremely limited similarities.

denoted by $omit_{p,p'}$, which takes the value of 1 if paper p cites paper p' in its references, and 0 otherwise:

$$omit_{p,p'} = \begin{cases} 1 \text{ if p does not cite p' conditional on p' in } \mathcal{P}_p \\ \\ 0 \text{ if p cites p' conditional on p' in } \mathcal{P}_p \end{cases}$$

This index therefore determines whether the relevant prior literature is included in the references of a given paper and to what extent.

4.5 Concrete example

Figure V shows a simple illustration of the construction of cosine similarity and the omission index. Consider three articles p, $p^{'}$ and $p^{''}$. Each article is represented by an abstract, which is a set of structured sentences. The first step illustrated by Panel (a) consists of writing each abstract as a set of words. We can further build a matrix which, for each word, indicates the number of times this word is used in a given article. In Panel (b), we calculate the relative frequency of each word in a given article (TF), and we multiply it by the IDF (which by definition depends on it of all the articles present in the database). Subsequently, the cosine between the articles p and $p^{'}$, for example, is defined by:

$$cos(p,p') = \frac{TfIdf(trade\_credit,p) * TfIdf(trade\_credit,p') + TfIdf(retailer,p) * TfIdf(retailer,p') + ...}{(TfIdf(trade\_credit,p)^2 + TfIdf(retailer,p)^2 + ...)^{1/2}(TfIdf(trade\_credit,p')^2 + TfIdf(retailer,p')^2 + ...)^{1/2}}$$

We do this for all the papers in the database. We obtain the cosine similarity matrix, which is a collection of pairwise similarity. We can now proceed to the construction of the omission index. In this example, p is the citing paper. $p^{'}$ and $p^{''}$ are in $\mathcal{P}_p$, the relevant prior literature for paper p. In other words, after ordering, the cosine pairwise involving p, p' and $p^{''}$ are among the highest. Panel (c) clearly indicates that p and p' have a lot in common; likewise for p and $p^{''}$. However, p did not cite $p^{'}$ and did cite $p^{''}$. Therefore, the omission index between p and $p^{'}$ is 1, $Omit_{p,p'} = 1$; and the omission index between p and $p^{''}$ is 0, $Omit_{p,p''} = 0$. Interestingly, $p^{'}$

is a female-authored paper, and p'' is a male authored-paper. This is the kind of pattern that the omission index captures on average.

4.6 Descriptives of cosine and omission index

Figure VI plots the empirical CDF of the pairwise similarity in panel (b) and the relationship between the pairwise similarity and the probability of being cited in panel (c). As demonstrated in panel (c), papers with high relative similarity are more likely to be linked by a citation. In other words, the probability that one paper cites another is increasing in the pairwise similarity between those two papers.

Figure VII, panel (a) reveals that about 55% of papers do not cite all of their top five most similar, and therefore most relevant, prior papers according to the distance metric. In other words, more than half of all published papers omit from their references at least one of the most relevant prior contributions to the literature. Panel (b) retains the number of related papers to $n = 5$, but imposes a minimum value on the cosine similarity. As can be seen, the results are pretty much the same. Panel (c) shows that the results are robust to the increase in the number of papers in the related prior literature, from $n = 5$ to $n = 10$. Finally, panel (d) shows that the results are unchanged when both the number of related papers and the degree of similarity across related papers are required to be higher.

Figure VIII presents the distribution of the number of papers cited from the relevant prior literature for a given paper. The horizontal axis is the number of related papers. The vertical axis shows the percentage of published papers that cite a given number of their most related papers. As can be seen, almost 60% of papers cite one of their most related prior papers, while 30% of papers cite two related prior publications. It is worth noting that in light of the huge economics literature, authors must choose between the articles they decide to cite and those they do not cite. However, this becomes problematic when this choice systematically excludes a category of people based on criteria such as gender. I investigate this case in the next section.

4.7 Omission and gender: Descriptive analysis

This section presents a descriptive investigation of the potential gendered pattern of the omissions of prior literature.

4.7.1 Gender and intersectionalities

Figure IX plots the evolution of the gender composition of authors of the most similar papers. Panel (a) shows that the fraction of papers whose most related literature, as determined by the pairwise similarity, includes at least one female author increased from 30% in the early 90s to 70% in 2018. However, the fraction of papers that completely omit any of those related papers increases from approximately 30% to 50% and the fraction of papers that cite some of those related papers only increases to 20%. Panel (b) shows that the fraction of papers whose most related literature includes majority-female authored papers increased from about 10% to more than 25% over the sample period. However, the fraction of papers that omit these related papers increases by more than 10 percentage points over the sample period. The fraction of papers that cite some of these related papers also increases to 5% in 2018. Essentially, despite the increasing and substantial representation of women authors in the related literature since 1991, the degree of their omission from references remains high over the past three decades.

Figure X presents the probability of omission as a function of gender and field of study. The field are grouped by their propensity to be more empirical versus less empirical. The odds of omission of papers written by women is on average 10% bigger in theoretical fields compared with applied fields. The gap between articles written by men and articles written by women is almost 4 times bigger in less empirical fields than in more empirical fields. In more empirical fields, relevant papers written by women are 5.66 times more likely of being omitted than cited, and relevant papers written by men are 5.16 times more likely of being omitted than cited. While the odds for men is 4.5 in theoretical fields, the odds for women worsen, with related papers written by women 6.21 times more likely of being omitted than cited. However, those numbers do not say anything about statistical significance. The model in the upcoming sections

will give more insights with this respect.

4.7.2 Gender distribution of omissions

An article may not refer to papers written by women or any other papers in a group of closely related literature, but instead could have a completely different gender structure in terms of the overall references it cites. This section aims to conduct a counterfactual analysis by looking at the gender structure of the references of all articles in the database. For example, suppose paper p is citing N other prior papers in the data. From this, we can obtain the observed distribution of references across authors' genders. Next, take the N most related papers to paper p, as determined by the pairwise similarity. From this, we can obtain the "target" distribution of references across authors' genders, i.e., the references that should have been included had similarity been the only determinant. Hence, actual citations are "accurate" when the observed distribution equals the target distribution of references. Figures XI and XII compare the gender distribution of the actual references with the target distribution of references.

Figure XI, panel (a), compares the actual distribution versus the target distribution. A positive difference means that the actual distribution of a certain gender type is higher than the target distribution of this gender type. A negative difference means that the actual distribution of a certain gender type is lower than the target distribution of this gender type. The distribution of the difference for females has a much greater weight in 0 compared with that of males. It also tends to be more oriented to the left, while the distribution of the difference for males tends to be more oriented to the right. For papers written by male authors, the average difference (actual-target) is about 3%, implying that papers written by male authors are more often cited compared with what is suggested by the target distribution. For papers written by women, the average difference is -1.5%, meaning that papers written by women are less often cited compared with the target distribution. In general, for a negative difference between the actual and target distributions, papers written by women are more likely to be totally omitted than cited to a lesser extent, compared with the target distribution. By contrast, papers written by men are more likely to be cited to a lesser extent than totally omitted from the references,

compared with the target distribution. Panel (b) complements this argument by plotting the distribution of the difference of both differences. The histogram has more weight on the right. This means that if males are *over-cited*, they are more *over-cited* than females. Similarly, when they are *under-cited*, they are less *under-cited* than females.

Figure XII plots the difference between the actual and target distributions over time for different genders. Roughly 75% of the papers exhibit a negative difference for females, i.e., they should cite more papers written by women than they actually do. This number is only half as large for papers written by males. Overall, there is a slight increase in the share of articles that have a positive difference (actual > target) for female citations, from almost 0 in the early 1990s to 10% in 2018. However, there is a noticeable increase in the share of articles that have a positive difference for male citations, from less than 10% in the early 1990s to almost 40% in 2018. The lack of citations of papers written by women is roughly constant over time. In other words, as shown in panel (a) of Figure XII, since 1991 there has been a constant decline in the accuracy of citations for papers written by men (green line). Instead, about 40% of publications (blue line) "over-cite" male papers (solo or all-male authors), meaning that the observed number of citations of male papers is higher than the target, which is based on similarity alone, and about 30% of publications (orange line) under-cite male papers. By contrast, as shown in panel (b), for papers written by women (solo or all-female teams), the accuracy remains roughly constant over the 1991-2019 period, with about 20% of published papers (green line) having accurate references. The fraction of publications that "over-cite" female papers is basically flat at zero (blue line), whereas the fraction of papers that under-cite women declines a bit over the period, but is still at the very high level of 80% (orange line).

5 Omission and gender: Empirical analysis

The previous section presented empirical evidence suggesting that gender potentially plays a role in determining the omission of prior related literature from a paper's references. This section performs a rigorous empirical analysis that controls for several factors that may influence the

observed omission patterns in economics publications.

5.1 Benchmark probability model

Assume that paper i was published in year t and paper j was published in year t'. Assume that paper j belongs in the relevant prior literature of paper i, according to the cosine similarity. Let $gender_j$ be the variable that defines the gender of paper j's authors. This is the variable of interest. To render the papers as similar as possible, the estimation equation includes a wide range of control variables. Let Z_j^1 and Z_i^2 be sets of controls for papers j and i, such as the journal of publication, authors' affiliation, number of authors, and number of references. Controlling for the journal is a way to condition on the quality of the paper. Similarly, the affiliations of the authors make it possible to exclude the presence of bias that could arise because women are less visible if they are more likely to be affiliated with lower-ranked institutions. The number of authors can also affect the probability of citing other authors because of the increasing size of the network as the number of authors grows. The number of an article's references makes it possible to exclude the possibility that the bias is systematic in articles with few references that will, therefore, choose only a handful of articles to cite. Let $Z_{i,j}^3$ be a set of controls for observed commonalities across papers i and j, such as the primary field of study. Let $Z_{t,t'}^4$ be a set of controls for the year of publication of each paper. The number of years between the cited and the citing papers can also affect the likelihood of a paper's being cited. The determinants of a paper's omission are investigated, given that this paper is similar to the citing one. Then, the probability of paper i omitting paper j, when it should have cited it according to the similarity distance, termed $omit_{ij}$, is given by the following logit model (or a corresponding linear probability model):

$$omit_{it,jt'} = \beta_0 + \beta_1 gender_j + \beta_2 Z_j^1 + \beta_3 Z_i^2 + \beta_4 Z_{i,j}^3 + \beta_5 Z_{t,t'}^4 + \epsilon_{it,jt'} \tag{6}$$

where standard errors are clustered at paper level.[13] The coefficient of interest is β_0 which captures the relative omission of female-authored papers (or any variable capturing the propensity of female authors in an article) compared with male-authored papers. In what follows, β_0 could be interpreted in percentages when the logarithm of the odds ratio is considered or in percentage points when the marginal probability is considered.

The results of this estimation are presented in Table *IV* for a number of different specifications and controls. The dependent or outcome variable is the probability of omission. It captures the probability that paper i cites paper j in the data, given that j is in the relevant prior literature of i, according to the similarity distance.

The variable "female" in the first row refers to an all-female author team (solo or multiple authors). The associated coefficient is the odds ratio that prior relevant paper j is omitted from the citations of paper i when the author team of paper j is all female compared with all male. On average, the coefficient is estimated between 20% and 30% and is always statistically significant at the 1% level. In other words, the odds of being omitted from the references are 20% to 30% higher for papers written by all-female teams compared with those by all male-teams.

The variable *Top 5 j* controls for whether a prior paper is published in a top-five economics journal. The associated coefficient, which is estimated between -0.5 and -0.7 across specifications, shows that the odds of being omitted from the references are on average 60% lower if the paper is published at a top-five journal rather than a non-top-five journal.

The variable *Primary field* denotes the same primary field of the citing and cited papers. Its coefficient is estimated between -0.5 and -0.6 and is statistically significant across all specifications. This shows that the odds of being omitted from the references are on average 55% lower if the paper belongs to the same primary field as the citing paper.

The variable *Years lag* captures the interval between publication years for citing and cited papers. The coefficient is estimated between 0.01 and 0.03 across specifications and is always

[13]We cluster standard errors at the citing paper level and/or at the cited/omitted paper level. Moreover, in all the tables the notation "R-sqr" designs either the pseudo R^2 or the adjusted R^2.

statistically significant. It shows that one additional year between publication dates increases the odds of being omitted by 2% on average.

The variable *Gender Structure* captures the gender structure of the relevant prior literature. Its coefficient is estimated between -0.14 and -0.17. This indicates that a one-unit increase in the share of papers with at least one female author in the relevant prior literature reduces the odds of being omitted by 15% on average. Put differently, a one-standard-deviation in the share of papers with at least one female author in the relevant prior literature decreases the odds of being omitted by 3%.

The variable *Number of Reference i* is the total number of references of a paper. Its coefficient is estimated around -0.05 across specifications and is statistically significant. This indicates that one additional bibliographical reference reduces the odds of being omitted by about 5%.

Overall, the results for the main variables of interest are significant and quantitatively similar across specifications. This finding is robust to the inclusion of fixed effects for the field, for the institutional affiliation of the authors of i and j, and for the journal and year of publication of paper i.

Column (1) of Table V shows the marginal probability of omission for all-female teams, which is 2.6 percentage points. The probability of being cited conditional on being in the most similar set is reduced by approximately 3 percentage points for female-authored papers compared with male-authored papers. This represents almost 15% of the mean of citation conditional on being in the most similar set of 18%. A simple back-of-the-envelope exercise reveals that for one citation of an all-male team, an all-female team will earn 0.84 citations.[14]

[14]Even if the study is not about wage, it is convenient to draw a parallel to the gender pay gap. Recent news coverage on ABC shows that women earned 84.7 cents for every dollar earned by their male counterparts in 2019: https://abcnews.go.com/Business/gender-pay-gap-persists-executive-level-study-finds/story?id=75945000. Goldin (2014) shows that a woman earns approximately 0.77 dollars for every dollar earned by a man. Freund et al. (2016) find, in a sample of faculty followed over 17 years, that women continued to earn 90 cents for every dollar a man earns.

5.2 Two-sided gender

This section examines in more detail the role of the gender structure of the citing and cited papers in the probability of omission. The dependent variable is the same as before. The controls now include a number of cross-gender variables:

$$omit_{it,jt'} = \tilde{\beta}_0 + \tilde{\beta}_1 gender_j + \tilde{\beta}_2 Z_j^1 + \tilde{\beta}_3 Z_i^2 + \tilde{\beta}_4 Z_{i,j}^3 + \tilde{\beta}_5 Z_{t,t'}^4 + \tilde{\beta}_6 \cdot gender_i + \tilde{\beta}_7 \cdot gender_i \cdot gender_j + \epsilon_{it,jt'} \tag{7}$$

In an ideal setting, the interaction effect reflects the difference-in-differences in the relative omission bias for a paper j written by an all-female team versus one j written by an all-male team, when paper i is written by all women relative to when paper i is written by all men.

The results are presented in Table V. The variable *female j* indicates only female authors in the cited paper (solo or all-female team). The variable *female i* indicates only female authors in the citing paper (solo or all-female team). The variable $A1f_j$ indicates at least one female author in the cited paper. The variable $A1f_i$ indicates at least one female author in the citing paper. The variables *female j* $\cdot$ *female i*, *female j* $\cdot$ $A1f_i$, $A1f_j$ $\cdot$ *female i* and $A1f_j$ $\cdot$ $A1f_i$ are cross-variables for citing and cited/omitted papers.

Let us consider column (2), in which the citing paper i has an all-female team, *female i* and paper j has an all-female team, *female j*. First, the coefficient on *female j* corresponds to $\tilde{\beta}_1 = 0.046$ and is statistically significant. This means that having only male authors in citing paper i increases the probability of being omitted for an all-female relevant paper j by 4.6 pp, compared with an all-male paper j. In other words, conditional on an all-male citing team, the probability of omission is 5 percentage points higher for female papers than for male papers.

Second, the coefficient on *female i* corresponds to $\tilde{\beta}_6 = 0.009$, which is not statistically significant. This means that for an all-male relevant paper j, the probability of being omitted is the same, regardless of whether citing team i is all female or all male.

Third, the coefficient on *female j* $\cdot$ *female i* corresponds to $\tilde{\beta}_7 = -0.103$ and is statistically significant. This means that having only female authors in citing paper i decreases the probability of being omitted for a relevant all-female paper j by around 10 percentage points

compared with an all-male paper i. In other words, a complete change in the gender structure of the authors of i from male to female is associated with a substantial, statistically and economically, increase of 10 percentage points in the probability of citation for relevant female-authored papers.

Column (4) presents the results for the case in which paper i has at least one female coauthor, $A1f_i$, paper j has at least one female coauthor, $A1f_j$, and the gender-interaction term is $A1f_i \cdot A1f_i$. The results and the corresponding interpretations are qualitatively similar to those in column (1). Column (5) presents the results of the gender structure (*female j*, $A1f_i$), with gender-interaction term *female i* $\cdot A1f_j$. Column (6) presents the results of the gender structure ($A1f_j$, *female i*), with gender-interaction term $A1f_j \cdot$ *female i*. Again, the results are qualitatively similar to those in column (1). On average, across all columns, $\tilde{\beta}_0 > 0$ and statistically significant, $\tilde{\beta}_6 < 0$ and statistically significant, and $\tilde{\beta}_5 = 0$, i.e., having a female in the citing paper reduces the omission of other female-authored papers.

5.3 Heterogeneous effects and robustness

5.3.1 Heterogeneous effects

Tables VI to VII report the estimates for different subgroups.

Top-five versus non top-five: Table VI shows that switching from only males to only females in i reduces the probability of being omitted for paper j published in a top-five (respectively, non-top-five) journal written by only women by 10 percentage points (respectively, 10 percentage points). The omission bias is present in both top-five and non-top-five journals. Thus, even women publishing in top-five journals are not exempt from omission bias.

Split by affiliation types: Furthermore, switching from only males to only females in i reduces the probability of being omitted for paper j from a top-tier affiliation written by only women by 9 percentage points. Overall, switching from only males to only females in paper j from a top-tier affiliation increases the probability of being omitted by 3 percentage points when i is written by only males. Similarly, switching from only males to only females in i reduces

the probability of being omitted for paper j from a mid-tier affiliation written by only women by 12 percentage points. Overall, switching from only males to only females in paper j from a mid-tier affiliation increases the probability of being omitted by 8 percentage points when i is written by only males. Finally, switching from only males to only females in i reduces the probability of being omitted for paper j from a low-tier affiliation written by only women by 10 percentage points. Overall, the omission of women relative to men tends to be bigger when comparing papers written by men and those written by women from mid-tier institutions.

Split by field: Table VII examines the robustness of the qualitative result $\tilde{\beta}_0 > 0$ and significant, $\tilde{\beta}_6 < 0$ and significant, across different fields in economics. As can be seen, the pattern is especially strong in columns (1) mathematical economics and econometrics; (2) microeconomics; (3) macroeconomics; (4) international economics; and (5) finance. Those fields show a higher probability of omission of relevant papers that include at least one female author when the citing team consists mostly of men compared with mostly women. By contrast, the omission of teams with at least one female author is not as strong in the fields of labor and education (column (6)) and industrial organization (column (7)). Overall, fields that are more theoretical and mathematical display a higher level of female-authored papers' omission when we have only male authors on the citing paper, while fields that are more empirical display a lower level of female-authored papers' omission when we have only male authors on the citing paper.

5.3.2 Robustness

Issues related to sample selection or bias in the algorithm could be raised as possible drivers of the estimated positive relation between omission and females. To verify that these elements do not lead the results, in this section I present a broad set of robustness checks.

Similarity level: Column (2) of Table VIII adds cosine similarity to the set of controls. The coefficient on cosine similarity is negative and statistically significant at -0.84. This means that a one-standard-deviation increase in the textual similarity across paper i and prior relevant paper j reduces the probability of omission of j by about 8.4 percentage points. Incidentally,

this confirms the use of cosine similarity as an index of contextual proximity across papers. Interestingly, the effect of an increase in the cosine is not uniformly distributed across gender: One additional unit in terms of the similarity between i and j reduces the probability that i omits j by 85% if j is written by only men and by only 60% if j if written by only women. In other words, a one-standard-deviation increase in the cosine similarity between i and j reduces the probability that i omits j by 8.5 percentage points if j is written by males and 6pp if j is written by females. The 2.5 percentage points gap is in line with the marginal probability of female omission of 2.6 percentage points. Therefore, even relatively high values of cosine similarity do not narrow the omission gap between males and females.

Legitimacy of the algorithm One concern that could be raised regards the legitimacy of the algorithm: it could be unclear how well the algorithm captures the similarity between papers. Because there is no metric to assess the accuracy of the algorithm (as is well known in the unsupervised machine learning literature), I use alternative methods to assess the reliability of the algorithm.

First, the validation in Section 4.6 indicates that high cosine values are associated with a higher probability of being cited (similarly, we could refer to column (2) of Table VIII). Then the cosine is capturing how likely a paper is to cite another one. Thus, the metric is suitable for capturing proximity between papers. I complement this by showing in the appendix an example of papers that are found to be similar in the data.

Second, I construct a set of placebo experiments in which I randomly simulate the most similar set to a given paper and conduct 1,000 simulations. Figure XIII shows the results of those experiments. In 22% of cases, the coefficient associated with the all-female-authored papers cannot be estimated. Conditional on being estimated, less than 2% of the estimated coefficients have a t-statistic lower than -1.96 (less than 3% for 1.64). The largest t-statistic is 1.18. This confirms that the algorithm is not capturing some random patterns in the data.

Third, even if there is a bias in the algorithm, it is unlikely that this is related to the citing article's characteristics: I found, with the two sided-gender effects, that having males or females in the citing paper influences how likely a female-authored paper is of being omitted.

The omission behavior is not random. Nevertheless, one could also argue that men's articles would be less similar because men would distinguish themselves by using a singular vocabulary. This issue could be viewed as the biased algorithm problem in which biased input data could produce biased output conclusions. To verify this hypothesis, I first consider the differences in IDF. Remember that the IDF captures the frequency of a word's use in the entire corpus. If men use infrequent words that make it harder to pick up on their similarity to other papers, they should have relatively high IDFs. Interestingly, columns (1) and (2) of Table IX show no difference between men and women in terms of the use of uncommon words. If anything, the results suggest the use of more uncommon words in favor of mixed-gender teams (for some metrics).[15]

Finally, I avoid reliance on a specific metric; to test whether my results are influenced by the choice of a given metric, I consider five alternative algorithms.[16] First, I use fully unsupervised cosine similarity. Second, I use a TFIDF commonly used in natural language processing that disregards language adjustment. Third, I implement soft cosine similarity, which allows a link between words (word embeddings). Fourth, I apply the Doc2vec algorithm, which goes one step further than soft cosine and incorporates sentences and paragraphs in embedding algorithms. Lastly, I construct a metric based on shared references between two articles.[17] More details on these variants can be found in the online appendix.

Columns (3) to (7) of Table IX show the result of those estimations. For each of these methods, the same qualitative results are obtained: There is gender bias in the omission of references. The minor difference is that the coefficient for representation (7) is relatively smaller.

[15]It is worth noting that the style of writing (complex grammatical form, length of sentences, etc.) is not a factor for the algorithm used, because it is based on matching words. Thus, it is not influenced by the writing style. In short, it is not how one speaks that matters but the words used to express what one says.

[16]We can draw a parallel to the bias-variance trade-off in the supervised learning literature. For a given paper (the citing paper), let us consider the share of female-authored papers omitted relative to male-authored papers. If there exists a share s^* that is the true share, an algorithm is a function $\hat{f}$ that tries to predict the true share s^*. In the bias-variance trade-off, a good algorithm has a low bias and a low variance. I have already shown that the algorithm is capturing patterns in the data, which supports a relatively low bias. The variance appears each time we change the training sample. In this case, a change in the training sample could come from a different way of capturing the similarity between papers. We expect a good algorithm to have a low variance, i.e., a change in the similarity metric should not change the gender-omission bias result.

[17]The reason the cross-references check is not a preferred metric is because it still includes the willingness of a given author to cite another one.

However, note that the latest algorithm (cross-reference) tends to perpetuate bias. Biased references will have more in common than unbiased references. Again, the value here is more likely to be a lower bound.

Other Robustness: Tables X and XI present other robustness checks.

In column (1) of Table X, like Angrist et al. (2017) and Card et al. (2020), I distinguish between empirical papers, theoretical papers, and structural papers. An argument could be made that men write in more theoretical fields. The similarities perhaps link women from the same literature but use empirical methods to men from the same literature who take a more theoretical approach, which is why the former are omitted. While it is common for an empirical paper to cite a theoretical paper and vice versa, it is also important to control for such a pattern to avoid the results relying entirely on this cross-methodological citation. In doing so, I use a pre-trained sample of around 1,000 articles to construct a classifier for the style of a given paper. Controlling for these characteristics does not change the results. Moreover, those columns confirm the previous intuition whereby omission tends to be lower in empirical fields than in theoretical fields.

Columns (3) and (4) analyze the effect over time. We can see that restricting the sample to 2000 and after or 2007 and after does not change the qualitative results, which show the relative omission of female-authored papers of around 20%. This reveals that time is not the driver of the gender omission bias.

Further, I analyzed whether the placement of the woman's name influences the omission bias. The results in column (5) show that the position of the woman's name does not affect the omission bias; the coefficient is estimated at 1.5% and is not significant. This result can be explained by the fact that in economics, the order in which the authors' names are listed does not relate to the relative productivity of the authors but rather the alphabetical order of the names.

Column (6) controls for the number of authors in the cited or omitted paper. The coefficient drops slightly to approximately 15%: More than 75% of papers written by only females are solo-authored papers. By contrast, 33% of papers written by only males are solo-authored papers.

The number of authors in the cited or omitted paper is correlated with the gender. However, this is not a sufficient argument to say that female-authored papers are omitted because women tend to write alone.

Table XI presents the results controlling for the share of females, for a battery of fixed effects, considering at most 10 articles in the most similar set, and extending the sample to more than 100 economic journals. The results are robust to all of these alternative specifications.[18]

6 Potential mechanisms and additional results

In this section, I study the relevance of some potential explanations for the gender omission bias. After that, I present additional evidence supporting my results.

6.1 Potential mechanism

Several explanations can be given for the previous results. In this section, I present some mechanisms that can be tested with the available data. I present evidence that gender omission bias is more than a problem of information or the relatively less experience of women in the field, but rather seems to be linked to the intrinsic fact of being perceived as a female.

6.1.1 Information

To test whether the omission is due to an information problem, I test the relative omission of women and men in a context in which there is an increased likelihood that authors know each other. I consider the case in which authors share the same institution. For example, if team i and team j both include a coauthor from the Harvard economics department, then a least one of the coauthors on papers i and j have the same affiliation. I consider another connectivity index that encompasses the fact of having the same affiliation, but also of being coauthors or having a shared coauthor. Overall, this captures the effect of an author's peers (peer effects).

[18] Additional robustness tests are presented in the online appendix. Those include, among others, controlling for self-citation.

Table XII examines the robustness of main qualitative results in the presence of peer effect.

Overall, having authors from the same institution reduces the probability of related paper j being omitted from the references of paper i by 7.5 percentage points if paper i is all-male authored and by approximately 5 percentage points if i is all-female authored. Similarly, the global connectivity index shows the same type of relation, whereby the benefit of a connection is reduced by almost 2 percentage points for female-authored papers. This suggests that peer effects are more beneficial for men than for women, even in cases in which both genders benefit from being in a close network.[19]

6.1.2 Androgynous name versus Non-androgynous name

Next, we compare the omission bias for women's names, perceived and known as such, to names that correspond to women in the database but can be used for both men and women. The goal is to see whether the observed bias is due to one paper's perception as a female paper. For example, consider first names like "Gray" or "Chen". Those names are used both for males and females. Therefore the probability of occurrences of those names in the population when the person is a woman could be 50%. By contrast, take first names like "Anna", "Maria". Those names tend to be attributed to women, with a probability near 90%. Let us say that we know (I actually manually check for those) that "Gray" and "Chen" are women in the database. What happens when we compare women with names "Gray" or "Chen", for which a priori we are not sure that those persons are women, with women named "Anna" and "Maria" for whom there a high consensual belief that they may be females? Technically speaking, the estimation focuses on a sample of female papers, distinguishing between names known as female with a higher probability and names known as female with a relatively lower likelihood. For example, we can compare the results for female names with a probability lower than 0.4, using names known as female with a probability greater than 0.4 as a reference. Therefore, the coefficient

[19] I additionally consider controlling for access to seminars and workshops for a given paper. To check this, I wanted to exploit the acknowledgement section on the title page of the paper. Unfortunately, in most cases (three papers out of four in a sample of 3,000 papers), authors state something like "We thank participants from various seminars and conferences" without listing the complete seminar or conference titles. An ongoing data collection by Doleac et al. (2021) could be helpful for future works in this direction.

must be read relative to the latter category.

Table XIII columns (1)-(4) present the results of the exercise described above, controlling for the same set of variables as in the baseline. The results are particularly striking.

The odds of omission are around 32%, significantly lower for female names with a probability below 0.4, using names known as female with a probability greater than 0.4 as a reference. This number reaches 48% when considering female names with a probability lower than 0.7, using names known as female with a likelihood greater than 0.7 as a reference. In other words, a paper's perception as a female paper is essential in driving the omission pattern. The more the article is perceived to be female, the more it will tend to be omitted. This result confirms that the mechanism supporting omission bias is indeed gender-related.

Similarly, we isolate typically non-white names (example: "Chen", "Deepu") from white names (example: "Anne","Peter"). Taking the WIPO name dictionary, we can associate each name with the country in which the name is used. To gain more power, we define European and American names as "typical white" names and consider the remaining as "non-typical white" names (note that the vast majority of this category consists of Asian names). The last column of Table XIII shows the results of this estimation. The omission odds are around 16% lower for female names that are non-typical white names compared with female names that are typical white names. Despite the fact that those results are insignificant and could hide a racial component, they tend to have the same qualitative content as the androgynous versus non-androgynous analysis.

6.1.3 Seniority and Experience

I approximate seniority and experience using different variables. The first aims to capture the time spent in the profession by the most senior author. This could also be viewed as a proxy for the age of the author in the profession. Thus, I consider her/his first publication year in the database and take the minimum at the paper level. Alternative seniority measures include the total number of publications of the most prolific author, the number of publications in top-five journals of the most prolific author, and being a superstar in the profession (e.g., an editor,

Nobel prize winner, etc.).

The controls related to experience and seniority do not change the qualitative results of this study (see Table XIV). There is still an omission bias, whereby the odds of omission of articles written by women are 15% to 20% greater than those of articles written by men. Although a slight reduction compared with the benchmark estimation, the results are still substantial in magnitude. For example, controlling for the age and number of publications, papers written by women have a 15.4% greater odds of omission than papers written by men. If the most senior authors on both articles have the same number of papers in top-five journals, papers written by women have a 17.7% greater odds of omission than papers written by men with a standard error of 0.043. Likewise, with the same number of total publications, papers written by women have 14.5% greater odds of omission than papers written by men with a standard error of 0.043. The reduction in the magnitude of the coefficients suggests that differences in terms of seniority and age in the profession explain part, but not all, of the omission of women relative to men.

To investigate how the effect of seniority depends on the gender of the article, I look at the differential effect of an additional publication by gender. I consider the number of publications up to the year of publication of the citing paper — in other words, how an additional publication in the publication history of the most prolific author influences the omission of a given paper. Ultimately, this will provide shreds of evidence of how the system updates information about an author's growing recognition depending on the gender.

Column (7) of Table XIV shows that having one additional publication reduces the probability of omission for men and women. However, the interaction term is insignificant, which means that there is no updating effect. Therefore, as the number of publications increases, the difference in omission patterns between female-authored papers and male-authored papers is not corrected.

6.1.4 Interpretation

The empirical results yield insights on the drivers of the gender omission bias. First, the analysis of peer effects and androgynous first names versus female first names clearly shows that the

results are not driven by a lack of information regarding articles written by women. Rather, they are related to the fact that women write these articles. Therefore, this rules out the assumption of lack of information. Second, the two-sided gender effect also shows that men and women have different preferences when citing women. Moreover, switching from men to women in the citing paper does not reduce men's omissions. Therefore, women's behavior is not particularly consistent with a homophilic pattern. The absence of a strong effect regarding the link between the history of publications and the female author's omissions impedes making a statistical discrimination interpretation (either correct or incorrect statistical discrimination, also called biased belief (Bohren et al. [2019]). Seniority matters, but seniority does not correct the observed difference between male-authored papers and female-authored papers. At the same time, the relatively lower omission bias in female-dominated subfields hides both (1) compliance: Females in the female-dominated fields are more likely be referee; (2) recognition pattern with a stereotyping behavior (Bordalo et al. [2016]): women in female-dominated fields could be expected to produce papers of high quality and therefore are less omitted.

Overall, junior-authored paper, especially junior female-authored papers are more likely to be omitted. Junior females are more likely to be omitted than senior females, but the latter are more omitted than their senior males colleagues. The more likely to omit are male-authored papers (either junior or senior).

6.2 Additional results

The previous sections have shown how strong and persistent the gender-omission pattern is. In the following, I provide evidence of gender-omission bias in other fields besides economics. I also show how omissions shape the relationship between citations and an alternative measure of a given paper's quality.

6.2.1 Comparison with other fields: economics vs mathematics vs sociology

To check whether gender bias in omission from references is a phenomenon peculiar to male-dominated fields, I compare economics with mathematics and sociology. I chose those fields

because traditionally, mathematics is more dominated by men than economics and sociology is less dominated by men than economics.[20]

I collect data from articles published in around 90 journals in mathematics and more than 200 journals in sociology between 1990 and 2019.[21] In the sample of mathematics papers, I have 116,466 papers, of which 3.69% of the papers are written by women. In the sample of sociology papers, I have 100,862 papers, of which 23.46% are written by women.[22]

Further, I construct the omission index for each pair of articles consisting of the article that cites and the article cited or omitted based on a similarity matrix specific to each field. I then estimate a benchmark equation similar to Equation 6 and control for field-specific effects, i.e., the fraction of papers published by women over time in each field.

Table XV shows that overall, there is omission bias in sociology and mathematics. However, compared with economics, the odds of women's omission relative to men's omission is 11% lower in sociology. On the other hand, the odds are 5.8% higher in mathematics but the coefficient is not significant. The results are in line with others in the literature, showing that gender bias is relatively less prominent in sociology than economics.

6.2.2 Citations and Innovativeness index

The literature suggests that citations are a noisy signal of quality; for example, in the case of patents. The analysis above also indicates that citations may not accurately reflect the quality of a published paper in economics, as they tend to systematically omit the contributions of female economists and groups of female economists. Therefore, in this section, following Kelly et al. (2018) and Koffi and Panousi (2021), I construct an alternative index to measure the quality of a publication in economics—the innovativeness index—which conducts a textual

[20]West et al. (2013), in a study using information from JSTOR, find that for the period 1990 to 2011, about 13.68% of the authors are women in economics, 10.64% in mathematics, and 41.44% in sociology.

[21]The full list of journals can be found in the online appendix. Many publications in sociology tend to be classified as book reviews. For example, for the American Sociological Review, one of the leading sociology journals, more than half of the publications listed in WoS are in the form of book reviews. This explains the large number of journals collected.

[22]The shares of papers written by only females in sociology and mathematics mimic, respectively, the shares found in Hunter and Leahey (2008) and Mauleón and Bordons (2012).

and linguistic comparison across different papers. By constructing a measure of quality that ignores the authors' willingness to cite articles (and is therefore without bias in this sense), we can assess the differential relationship between men and women by comparing this measure of quality with no inherent gender bias and citations in which gender bias has been highlighted in the previous section.

Specifically, the new quality index, denoted by q, has two dimensions that together capture the degree of innovativeness of a paper. First, more innovative papers are more distinct from prior related papers, in that they contribute a novel idea or method to the preexisting stock of knowledge. Second, more innovative papers are more likely to influence the framework or the methodology of future papers. In other words, the concept of innovation used here reflects the novelty as well as the influence of a publication. Papers with a high innovativeness index are both novel (distinct from prior papers) and influential (similar to future papers). Overall, the most important or significant papers introduce new concepts to the literature that render them different from their predecessor but very useful for future advances in economics. The "novelty" of a paper is captured by a backward-similarity index, which is the sum of pairwise relative cosine similarities of paper p, published in t, to paper p', published in $t - T$:

$$BS^0_{-T}(p) = \Sigma_{p'} \tilde{\lambda}_{p,p'}$$

Papers with low backward similarity are dissimilar from the past literature, and hence they are innovative compared with the existing literature.

The "influence" of a paper is captured by a forward-similarity index. The forward similarity is the sum of pairwise cosine similarities of paper p, published in t, to paper p', published in $t + T$:

$$FS^T_0(p) = \Sigma_{p'} \tilde{\lambda}_{p,p'}$$

Papers with high forward similarity have a higher impact on future publications. For example, they may open up a rich future line of literature, or they may propose an empirical methodology

that many future papers will use.

Therefore, the innovativeness index q will be a combination of the novelty and the impact of a paper, as measured, respectively, by backward and forward similarity.

$$q^T(p) = \frac{FS_0^T(p)}{BS_{-T}^0(p)}$$

In the benchmark specifications, $T = 5$, but the results are robust to alternative windows.

The q-index is a measure of the underlying scientific innovativeness of a paper. If a paper has high forward similarity (high numerator) and high backward similarity (high denominator), this could mean that the paper is a follower among other followers in a research area. Hence, it will have a low q-index, compared with a paper with high forward similarity and relatively low backward similarity. In that respect, it operates like the citations measure. In the baseline, the horizon chosen to compute the q-index is 5 years.

Innovativeness, citations and gender This section examines the relationship between the number of citations, the innovativeness index, and the gender of the paper.

$$C_{pt} = a_1 \cdot Q_{pt} + a_2 \cdot gend_{pt} + a_3 \cdot gend_{pt} \cdot Q_{pt} + a_4 \cdot Z_p + \theta_t + \epsilon_{pt} \tag{8}$$

C_{pt} is the logarithm of the number of citations of paper p published in year t; Q_{pt} is the innovativeness index of paper p published in year t (I keep the five-year horizon for C and Q to have the same scale of comparison); $\hat{Z}_p^1$ is a set of paper-level controls, such as the number of coauthors, coauthors' affiliations, field of the paper, journal of publication, and NBER membership; $\hat{Z}_t^2$ captures publication-year fixed effects. Field and journal fixed effects are included. standard errors are clustered by publication year and journal. The variable *gend* takes the value 1 if there is at least one female author and 0 if the paper is written by all-male teams; or a dummy variable that takes the value 1 if the paper is written by only women and 0 if the paper is written all-male teams (solo and coauthored).

Overall, there is a positive and statistically significant correlation between the innovativeness index and the number of citations received by an article. However, there are strong heterogeneous effects depending on the gender of the authors on the paper, as shown in Figure XV. Plot (a) shows that for the same innovativeness index value, the male paper will get more citations than the female paper. For the same number of citations (let us say the mean value of citation), the innovativeness index is near 0.8 (close to the 50^{th} percentile) for male papers and near 0.85 (close to the 75^{th} percentile) for female papers . Panel (b) focuses only on top-five publications. For the same number of citations at the mean, male-authored papers are at the 70^{th} percentile of the quality distribution and female-authored papers are at the 90^{th} percentile. Hence the gap is persistent even when considering only the top-five publications.

Counterfactual analysis: Compensating citations with omissions Panel (c) of Figure XV is a crucial result, showing the effect of the bias generated by omissions. I conduct a counterfactual analysis, in which the total number of omissions is added to the number of citations. This is interpreted as the number of citations an article would have received if all of the papers with which it shares the most similarities had cited it. The gap corresponding to more than the 20^{th} percentile, in terms of the innovativeness index, disappears completely with this compensation. In other words, if without the omission bias, the standards for being cited would have been the same for women and men.

7 Gender omission bias and future productivity

This section examines the relationship between the history of past omissions and the future productivity of a given author. The analysis uses four measures of future productivity: (1) the probability of being published in a top-five journal within the next 3 years; (2) the total number of forward citations of all papers by the same author over the next 5 years; (3) the quality of future publications by the author over the next 5 years; and (4) the probability of having only one publication in the sample. To do this, I construct a database at author level that lists, for

each author, all of his/her publications.

The regression specification for the case in which the measure of future productivity is the probability of publication in the top five is:

$$Top5_{p,t,a} = \theta_1 \cdot H\_Omission_{p,t,a} + \theta_2 \cdot gender_a + \theta_3 \cdot \Gamma_a^1 + \theta_4 \cdot \Gamma_p^2 + \Gamma_t^2 + \varepsilon_{p,t,a} \tag{9}$$

Here, $Top5_{p,t,a}$ is a binary variable that indicates whether paper p by coauthor a is published in a top-five journal in year t. Next, $H\_Omission_{p,t,a}$ is the omission history of author a prior to paper p. This captures all of the previous times any paper written by author a belonged in the relevant literature, according to the similarity distance, but was omitted from the references. The control variables include author-specific variables such as gender and characteristics of past publications such as quality, citations, and number of top-five publications. Paper characteristics such as field, year of publication, and time lag between two publications are also included.

Table XVI presents the results. As can be seen in the first row of the table, an increase in the history of omissions is associated with a reduction of the probability of being published in the top five within the next 3 years. The coefficient is estimated at 2.8 percentage points. In other words, a one-unit increase in total past omissions reduces the probability of being published in a top-five journal by approximately 3 percentage points. Concretely, a one-standard-deviation increase in past omissions measured with the inverse hyperbolic sine, which corresponds to being omitted by approximately two-three articles, is associated with a 5 percentage points reduction in the probability of being published in a top-five journal (15% of the mean probability of being published in a top-five journal in the database). Similarly if we take the correspondence of one male citation for 0.84 female citation, women miss almost 2 citations per 10 citations received by males. Therefore, reducing their chance to be published by more than 2.8 percentage points (almost 10% of the mean probability of being published in a top-five journal in the database). This effect seems sufficiently large to impact career path.

Column (2) shows similar results for the number of future citations. Again, a one-unit

increase in past omissions is associated with a decrease in the future paper's citations by 7 log points. Similarly, in column (3), a one-unit increase in past omissions is associated with a decrease in the future paper's innovativeness index by 0.003 units. Lastly, columns (4) shows the link between past omissions and the probability of having only one publication in the sample ("one shot"). I define a "one-shot" event as the probability of having only one publication in the sample.[23] For column (4), a one-unit increase in average omissions is associated with a 3.5 percentage points higher likelihood of having a one-shot event. In other words, two more papers that omit a given article increases the chance of the author of that article having only one publication and exiting the sample (and perhaps the field) by approximately 5 percentage points (roughly 10% of the mean of one shot).

Overall, despite the suggestive nature of the underlying results, there seems to be a link between past omissions and future productivity, with a reduction of the probability of being published in the top five journals over a 3-year horizon, a reduction in the future number of citations, a reduction in the innovativeness index, and an increase in the probability of exiting the sample.[24]

8 Discussion and Conclusion

Women are still underrepresented in math-intensive fields. However, few studies have sought to analyze whether the potential problem lies in a lack of recognition of their work. Accordingly, this paper examines the issue using data on economics and demonstrates that women have a higher probability of being omitted from references. This problem is persistent, even for women publishing in top journals in the same manner as men. However, the most vulnerable population appears to be women at mid-tier institutions; the bias is lower when comparing

[23] I also alternatively defined "one-shot" event as the probability of having only one publication in the sample and the single publication is dated before 2010. This does not change the results. Restricting to a date far in the past increases the chance of capturing true single publications and possibly dropping out of the sample (of the field) instead of simply the difficulty of publishing in top journals.

[24] The lower differential effect by gender could be due to the fact that to assess the link to publication in top-five journals, citations, and the innovativeness index, the author must have at least two publications. Therefore, there is sample selection.

female-authored papers from top-tier institutions with similar male-authored papers.

But what drives the gender-omission pattern? I tested several explanations related mainly to the lack of information, the network, and the lack of seniority of females. None of them explained the gap between male-authored papers and female-authored papers in terms of omission from references.

Overall, the results of this paper have strong implications on the policy side. By highlighting this previously undocumented phenomenon, it raises scholars' awareness of the issue. For example, editors and referees may pay more attention to the omission of female-authored references when evaluating a paper. Institutions may take this relative under-citation of female-authored papers into account in making tenure decisions or awarding grants; the omission index could constitute a barometer in many instances.

Finally, beyond documenting a gender omission bias in economics, this study aims at a more general horizon by explaining how discriminating factors can influence the perception of individual works even when they are more deserving than others. Besides recognizing how inequality issues can affect socioeconomic factors, this paper advocates for greater inclusion of minorities to increase overall productivity.

References

Joshua Angrist, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–97, 2017.

Heather Antecol and Deborah A Cobb-Clark. Do psychosocial traits help explain gender segregation in young people's occupations? *Labour Economics*, 21:59–73, 2013.

Heather Antecol, Kelly Bedard, and Jenna Stearns. Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? *American Economic Review*, 108(9):2420–41, 2018.

Emmanuelle Auriol, Guido Friebel, and Sascha Wilhelm. Women in european economics. 2019.

Amanda Bayer and Cecilia Elena Rouse. Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives*, 30(4):221–42, 2016.

Francine D Blau and Jed DeVaro. New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires. *Industrial Relations: A Journal of Economy and Society*, 46(3):511–550, 2007.

J Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436, 2019.

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794, 2016.

Lutz Bornmann, Alexander Butz, and Klaus Wohlrabe. What are the top five journals in economics? a new meta-ranking. *Applied Economics*, 50(6):659–675, 2018.

David Card and Stefano DellaVigna. Nine facts about top journals in economics. *Journal of Economic Literature*, 51(1):144–61, 2013.

David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberri. Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics*, 135(1):269–327, 2020.

Stephen J Ceci, Donna K Ginther, Shulamit Kahn, and Wendy M Williams. Women in academic science: A changing landscape. *Psychological science in the public interest*, 15(3):75–141, 2014.

Anusha Chari and Paul Goldsmith-Pinkham. Gender representation in economics across topics and time: Evidence from the nber summer institute. Technical report, National Bureau of Economic Research, 2017.

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

Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. Gendered citation patterns across political science and social science methodology fields. *Political Analysis*, 26 (3):312–327, 2018.

Carlo D'Ippoliti. "many-citedness": Citations measure more than just scientific quality. *Journal of Economic Surveys*, 2021.

Jennifer L Doleac, Erin Hengel, and Elizabeth Pancotti. Diversity in economics seminars: who gives invited talks? In *AEA Papers and Proceedings*, volume 111, pages 55–59, 2021.

Lorenzo Ductor, Sanjeev Goyal, and Anja Prummer. Gender and collaboration. *DP15673*, 2021.

Glenn Ellison. How does the market use citation data? the hirsch index in economics. *American Economic Journal: Applied Economics*, 5(3):63–90, 2013.

Kristie M Engemann, Howard J Wall, et al. A journal ranking for the ambitious economist. *Federal Reserve Bank of St. Louis Review*, 91(3):127–139, 2009.

Marianne A Ferber. Citations: Are they an objective measure of scholarly merit? *Signs: Journal of Women in Culture and Society*, 11(2):381–389, 1986.

Marianne A Ferber. Citations and networking. *Gender & Society*, 2(1):82–89, 1988.

Eric A Fong and Allen W Wilhite. Authorship and citation manipulation in academic research. *PloS one*, 12(12):e0187394, 2017.

Karen M Freund, Anita Raj, Samantha E Kaplan, Norma Terrin, Janis L Breeze, Tracy H Urech, and Phyllis L Carr. Inequities in academic compensation by gender: a follow-up to the national faculty survey cohort study. *Academic medicine: journal of the Association of American Medical Colleges*, 91(8):1068, 2016.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57(3):535–74, 2019.

John Gibson, David L Anderson, and John Tressler. Citations or journal quality: which is rewarded more in the academic labor market? *Economic Inquiry*, 55(4):1945–1965, 2017.

Donna K Ginther and Shulamit Kahn. Women in economics: moving up or falling off the academic career ladder? *Journal of Economic perspectives*, 18(3):193–214, 2004.

Claudia Goldin. A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119, 2014.

Daniel S Hamermesh. Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature*, 56(1):115–56, 2018.

Daniel S Hamermesh and Gerard A Pfann. Reputation and earnings: the roles of quality and quantity in academe. *Economic Inquiry*, 50(1):1–16, 2012.

James J Heckman and Sidharth Moktan. Publishing and promotion in economics: the tyranny of the top five. *Journal of Economic Literature*, 58(2):419–70, 2020.

Erin Hengel. Publishing while female (summary). 2020.

Erin Hengel and Euyoung Moon. Gender and equality at top economics journals. 2020.

Michael J Hilmer, Michael R Ransom, and Christiana E Hilmer. Fame and the fortune of academic economists: How the market rewards influential research in economics. *Southern Economic Journal*, 82(2):430–452, 2015.

Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291, 2020.

Laura Hospido and Carlos Sanz. Gender gaps in the evaluation of research: Evidence from submissions to economics conferences. *Oxford Bulletin of Economics and Statistics*, 83(3): 590–618, 2021.

Jennifer Hunt. Why do women leave science and engineering? *ILR Review*, 69(1):199–226, 2016.

Laura Hunter and Erin Leahey. Collaborative research in sociology: Trends and contributing factors. *The American Sociologist*, 39(4):290–306, 2008.

Pablo Jensen, Jean-Baptiste Rouquier, and Yves Croissant. Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78(3):467–479, 2009.

Pantelis Kalaitzidakis, Theofanis P Mamuneas, and Thanasis Stengos. Rankings of academic journals and institutions in economics. *Journal of the european economic association*, 1(6): 1346–1366, 2003.

Pantelis Kalaitzidakis, Theofanis P Mamuneas, and Thanasis Stengos. An updated ranking of academic journals in economics. *Canadian Journal of Economics/Revue canadienne d'économique*, 44(4):1525–1538, 2011.

Graciela L Kaminsky and Carmen M Reinhart. The twin crises: the causes of banking and balance-of-payments problems. *American economic review*, 89(3):473–500, 1999.

Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. Measuring technological innovation over the long run. Technical report, National Bureau of Economic Research, 2018.

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

Marlène Koffi. Gendered citations at top economic journals. In *AEA Papers and Proceedings*, volume 111, pages 60–64, 2021.

Marlène Koffi and Vasia Panousi. Patents, innovation and growth in canadian pharmaceuticals. *Unpublished working paper*, 2021.

David N Laband and Michael J Piette. The relative impacts of economics journals: 1970-1990. *Journal of economic Literature*, 32(2):640–666, 1994.

Ryan Lampe. Strategic citation. *Review of Economics and Statistics*, 94(1):320–333, 2012.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

Shelly Lundberg and Jenna Stearns. Women in economics: Stalled progress. *Journal of Economic Perspectives*, 33(1):3–22, 2019.

Elba Mauleón and María Bordons. Authors and editors in mathematics journals: A gender perspective. *International Journal of Gender, Science and Technology*, 4(3):267–293, 2012.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the national academy of sciences*, 109(41):16474–16479, 2012.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Anne E Preston. Why have all the women gone? a study of exit of women from the science and engineering professions. *The American Economic Review*, 84(5):1446–1462, 1994.

Margaret W Rossiter. The matthew matilda effect in science. *Social studies of science*, 23(2): 325–341, 1993.

Heather Sarsons. Recognition for group work: Gender differences in academia. *American Economic Review*, 107(5):141–45, 2017.

Heather Sarsons, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):000–000, 2021.

Dawn Langan Teele and Kathleen Thelen. Gender in the journals: Publication patterns in political science. *PS: Political Science & Politics*, 50(2):433–447, 2017.

Jevin D West, Jennifer Jacquet, Molly M King, Shelley J Correll, and Carl T Bergstrom. The role of gender in scholarly authorship. *PloS one*, 8(7):e66212, 2013.

Allen W Wilhite and Eric A Fong. Coercive citation in academic publishing. *Science*, 335 (6068):542–543, 2012.

Alice H Wu. Gendered language on the economics job market rumors forum. In *AEA Papers and Proceedings*, volume 108, pages 175–79, 2018.

Basit Zafar. College major choice and the gender gap. *Journal of Human Resources*, 48(3): 545–595, 2013.

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427, 2015.

Tables

Table I:
Summary statistics journal level

	Full sample	Male	Female	Mixed	Undetermined
	(1)	(2)	(3)	(4)	(5)
			Overall		
Total Number	24,033	17,785	1,091	4,255	902
			Journal		
American Economic Review	1931	1,398	108	378	47
Econometrica	1,485	1,244	39	186	16
Journal of Econometrics	3,089	2,230	96	532	231
Journal of economic growth	260	193	10	47	10
Journal of economic literature	340	266	26	41	7
Journal of economic perspectives	1,026	809	60	144	13
Journal of Economic Theory	2,651	2,138	70	312	131
Journal of Finance	2,001	1,509	60	395	37
Journal of Financial Economics	2,247	1,533	87	517	110
Journal of International Economics	1,775	1,146	167	371	91
Journal of Labour Economics	836	513	87	193	43
Journal of Monetary Economics	1,193	881	74	198	40
Journal of Political Economy	1,122	889	49	163	21
Quarterly Journal of Economics	1,105	830	49	213	13
Review of Economics Studies	1,201	936	56	183	26
Review of Financial Economics	1,771	1,270	53	382	66

This table presents the journals in the database and the total number of papers per journal over the sample period 1991-2019. The selected papers exclude proceedings, comments, articles of less than three pages, books reviews, bibliographical items, articles without references and without abstracts, editorial material, letters and corrections.

Table II:
Summary statistics publication level

	Full sample	Male	Female	Mixed	Undetermined
	(1)	(2)	(3)	(4)	(5)
Total	24,033	17,785	1,091	4,255	902
			Authors		
Single authored	6,949	5,950	836	0	163
Coauthored	17,084	11,835	255	4,255	739
			Field		
Mathematical	4,259	3,217	128	664	250
Microeconomics	4,051	3,243	123	536	149
Macroeconomics	1,938	1,514	79	305	40
International Economics	2,235	1,544	191	423	77
Finance	6,184	4,467	219	1,284	214
Labour & Education	2,514	1,645	211	577	81
IO	236	182	16	28	10
Other	2,616	1,973	124	438	81
			Institutions		
Top tier	9,019	6,743	345	1,741	190
Middle tier	7,238	5,164	299	1,468	307
Low tier	4,486	5,715	294	954	334
Undefined	2,061	1,745	153	92	71
			References		
Average Number	38.04	37.31	38.87	41.29	36.17
Average Number (database)	9.25	8.83	8.72	11.18	9.08

The table describes the papers published per journal in the database over the period 1991-2019. The field selection is based on the Journal of Economic Literature (JEL) codes. The category *Other* includes public economics, agricultural economics, general economics, urban economics, law and economics, business administration, economic history, and economics systems.

Table III:
Distribution of the cosine

Mean	0.24
Standard deviation	0.12
1th percentile	0.10
5th percentile	0.12
10th percentile	0.13
25th percentile	0.16
Median	0.21
75th percentile	0.28
90th percentile	0.39
95th percentile	0.48
99th percentile	0.70

The table shows the distribution of the relative cosine. For paper p and p_{max} such as: $p_{max} = \max_{p' \in C, p' \neq p} cos(p, p')$, for any given article p' in the database, the relative cosine of paper p and paper p' is defined as: $\tilde{\lambda}_{p,p'} = \frac{\lambda_{p,p'}}{\lambda_{p,p_{max}}}$.

Table IV:
Relationship between omission and gender

	Outcome variable: Omission				
	(1)	(2)	(3)	(4)	(5)
female j	0.289***	0.253***	0.257***	0.215***	0.215***
	(0.041)	(0.042)	(0.043)	(0.043)	(0.081)
Top 5 j		-0.497***	-0.558***	-0.692***	-0.692***
		(0.018)	(0.019)	(0.020)	(0.032)
Primary field		-0.586***	-0.564***	-0.503***	-0.503***
		(0.019)	(0.019)	(0.019)	(0.024)
Years lag		0.008***	0.023***	0.028***	0.028***
		(0.001)	(0.002)	(0.002)	(0.003)
Gender Structure			-0.142***	-0.168***	-0.168***
			(0.043)	(0.045)	(0.049)
Number of Reference i			-0.050***	-0.048***	-0.048***
			(0.001)	(0.002)	(0.002)
Number of Authors i				-0.028***	-0.028***
				(0.011)	(0.010)
Institution of j FE			Y	Y	Y
Institution, Journal, Field of i FE				Y	Y
Year of publication of i FE				Y	Y
N	110,767	110,767	110767	110,763	110,763
R-sqr	0.002	0.021	0.064	0.075	0.075

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i (as defined by 5). *female j* represents papers written by only women and is the variable of interest. The reference variable is *male*, which represents papers written by only men (the two other gender structure -Mixed and undetermined- are added but not shown in the table to ease the reading). The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by the citing papers, except in column (5) where the cluster is at the cited/omitted paper level. Standard errors are reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Table V:
Omission and two-sided gender

	Outcome variable: Omission					
	(1)	(2)	(3)	(4)	(5)	(6)
female j	0.026***	0.046***			0.046***	
	(0.005)	(0.006)			(0.006)	
female i		0.009				0.011
		(0.007)				(0.007)
female j · female i		-0.103***				
		(0.021)				
$A1f_j$			0.011***	0.031***		0.032***
			(0.003)	(0.003)		(0.003)
$A1f_i$				0.016***	0.015***	
				(0.003)	(0.003)	
$A1f_j \cdot A1f_i$				-0.070***		
				(0.007)		
female j ·$A1f_i$					-0.066***	
					(0.011)	
$A1f_j$· female i						-0.060***
						(0.013)
N	92,105	72,546	107,301	103,377	88,759	83,598
R-sqr	0.067	0.069	0.069	0.069	0.068	0.069

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper emphasizing the gender of the citing paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 5. *female_x* represents paper x written by only women. $A1f_x$ represents paper x with at least one female author. All the specifications include the full set of controls of the baseline model. The equations are estimated using a linear probability model. The size of the sample varies because of the selection in the specification considered. For example, column (2) includes only citing papers and cited papers that are written only by females or only males. The table displays the marginal probabilities. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table VI:
Omission and gender: by journals and institutions

	Outcome variable: Omission				
	Journal j		Institution j		
	(1)	(2)	(3)	(4)	(5)
	Top5	Non Top5	Top tier	Mid tier	Low tier
female j	0.053***	0.039***	0.028**	0.080***	0.029**
	(0.011)	(0.006)	(0.014)	(0.011)	(0.013)
female i	0.013	0.013*	0.031***	0.001	0.024*
	(0.012)	(0.008)	(0.012)	(0.013)	(0.012)
female j · female j	-0.101***	-0.099***	-0.090**	-0.123***	-0.098**
	(0.039)	(0.024)	(0.040)	(0.040)	(0.044)
N	25,433	47,113	24,543	18,306	13,370
R-sqr	0.130	0.094	0.117	0.094	0.079

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper emphasizing the gender of the citing paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 5. *female i* represents paper x written by only women. All the specifications include controls for paper j published in a top 5 journal; paper i and paper j having the same primary field; difference between the publication year of paper i and the publication year of paper j; the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a linear probability model. The table displays the marginal probabilities. The total number of observations in the institution section does not add up to 72546 (total number of observations in two-sided case with only females in the citing and the cited/omitted) because some affiliations are missing in the database. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table VII:
Omission and gender: field of study

	Outcome variable: Omission							
	(1) Mathe-matical	(2) Micro	(3) Macro	(4) International Economics	(5) Finance	(6) Labour - Education	(7) IO	(8) Other fields
$A1f_j$	0.030***	0.035***	0.043***	0.040***	0.023***	0.000	0.014	0.029***
	(0.008)	(0.009)	(0.011)	(0.010)	(0.007)	(0.010)	(0.030)	(0.009)
$A1f_i$	0.017**	0.017*	-0.005	0.024**	0.020***	0.008	-0.001	0.012
	(0.008)	(0.009)	(0.012)	(0.010)	(0.007)	(0.009)	(0.036)	(0.010)
$A1f_j \cdot A1f_i$	-0.092***	-0.082***	-0.072***	-0.075***	-0.078***	-0.011	-0.004	-0.034*
	(0.018)	(0.020)	(0.025)	(0.018)	(0.014)	(0.016)	(0.065)	(0.019)
N	17,519	17,119	8,549	9,851	27,881	10,858	998	10,602
R-sqr	0.135	0.107	0.116	0.127	0.089	0.115	0.098	0.117

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper emphasizing the gender of the citing paper and splitting by primary field of the citing paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 5. $A1f_x$ represents paper x with at least one female author. All the specifications include controls for paper j published in a top 5 journal; the relative cosine; paper i and paper j having the same primary field; difference between the publication year of paper i and the publication year of paper j; the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The field section is defined based on the Journal of Economic Literature (JEL) codes. The category "other fields" includes public economics, agricultural economics, general economics, urban economics, law and economics, business administration, economic history, economics systems. The equations are estimated using a linear probability model. The table displays the marginal probabilities. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table VIII:
Omission and gender: effect of cosine

	Outcome variable: Omission		
	(1)	(2)	(3)
female j	0.046***	0.044***	-0.027***
	(0.006)	(0.006)	(0.010)
female i	0.009	0.013**	
	(0.007)	(0.007)	
female j · female i	-0.103***	-0.099***	
	(0.021)	(0.021)	
cosine_ij		-0.839***	-0.854***
		(0.018)	(0.016)
female j · cosine_ij			0.237***
			(0.050)
N	72,546	72,546	92,105
R-sqr	0.069	0.108	0.107

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper emphasizing the effect of the similarity between two papers. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 5. *female j* represents paper x written by only women. *cosine_ij* is the value of the cosine between i and j. All the specifications include controls for paper j published in a top 5 journal; paper i and paper j having the same primary field; difference between the publication year of paper i and the publication year of paper j; the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a linear probability model. The table displays the marginal probabilities. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table IX:
Omission and gender: biased algorithm?

	Outcome variable: Omission						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Max IDF	Mean cosine	Unsupervised adjusted cosine	Basic Cosine	Soft Cosine	Doc2vec	Cross-Reference
female	-0.004	-0.007	0.131***	0.145***	0.234***	0.174***	0.101***
	(0.004)	(0.014)	(0.045)	(0.046)	(0.048)	(0.051)	(0.033)
mixed	0.005**	-0.005	-0.035	-0.032	-0.042	-0.030	0.040**
	(0.002)	(0.007)	(0.026)	(0.026)	(0.027)	(0.029)	(0.017)
Number of authors	Y	Y	Y	Y	Y	Y	Y
NBER	Y	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y	Y
Journal FE	Y	Y	Y	Y	Y	Y	Y
Affiliation FE	Y	Y	Y	Y	Y	Y	Y
Field FE	Y	Y	Y	Y	Y	Y	Y
Top5 j			Y	Y	Y	Y	Y
Same primary field			Y	Y	Y	Y	Y
Year difference		Y	Y	Y	Y	Y	
Gender structure similar set			Y	Y	Y	Y	Y
Number Reference			Y	Y	Y	Y	Y
N	24,033	24,033	113,247	114,309	115,873	115,039	182,423
R-sqr	0.036	0.142	0.100	0.113	0.080	0.136	0.058

This table presents the robustness checks associated with the algorithm. The key variable of interest is the female variable which corresponds to articles written by women. The results are in reference to articles written by men. Columns (1) and (2) are linear regressions where the dependent variables are respectively the maximum IDF and the mean cosine by papers. Columns (3) to (7) refer to the model of equation 6 where I present the relation between the omission and the gender of the authors of the article. Therefore, for Columns (3) to (7), the equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. The sample size vary because of the specificities of each similarity algorithm. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Table X:
Omission and gender: other robustness (1/2)

	Outcome variable: Omission					
	Methodology		Over time		Names' order	Number of Authors
	(1)	(2)	(3)	(4)	(5)	(6)
			> 2000	> 2007		
female j	0.212***	0.217***	0.191***	0.207***	0.168***	0.142***
	(0.043)	(0.043)	(0.046)	(0.055)	(0.051)	(0.044)
empirical	-0.062***					
	(0.021)					
structural	0.528***					
	(0.032)					
Same Methodology		-0.111***				
		(0.019)				
Male j					-0.080**	
					(0.032)	
Mixed team with Female at first place j					0.015	
					(0.044)	
Share of female						
Number Authors j						-0.112***
						(0.012)
N	110,763	110,763	87,467	59,537	110,763	110,763
R-sqr	0.079	0.076	0.061	0.065	0.075	0.074

This table shows additional robustness related to the baseline model. All the equations contain the same controls as in the baseline (Column (4) table IV). Column (1) and (2) add controls for the methodology (theory, empiric, structural and experimental). Column (3) and (4) present the change over time focusing respectively on citing papers published after 2000 and citing papers published after 2007. Column (5) presents the result putting as the reference group, papers written by mixed team with female not listed first in the authors names list. Column (6) presents the result controlling for the number of authors in the cited or omitted paper. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Table XI:
Omission and gender: other robustness (2/2)

	Outcome variable: Omission					
	Share of female by field		Additional Controls		10 most similar	Extented Sample
	(1)	(2)	(3)	(4)	(5)	(6)
	Primary Field	3 digit Jel				
female	0.215***	0.224***	0.217***	0.151***	0.230***	0.137***
	(0.043)	(0.043)	(0.044)	(0.045)	(0.035)	(0.023)
Share of female	-0.485	-0.576***				
	(0.422)	(0.079)				
Field · Year			Yes			
Journal · Year				Yes		
N	110,763	110,763	110,584	110,542	209,705	750,509
R-sqr	0.075	0.076	0.079	0.087	0.072	0.044

This table shows additional robustness related to the baseline model. All the equations contain the same controls as in the baseline (Column (4) table IV). Column (1) and (2) present the results of the estimation of equation 6 with the same controls, adding controls for the share of females in the subfield of the citing paper. Column (3) and (4) present the baseline adding respectively field and year fixed effects and Journal and year fixed effects. Column (5) presents the results of the estimation of equation 6 with the same controls, using 10 papers in the most similar set. Column (6) presents the results of the estimation of equation 6 with the same controls, extended the sample to around 100 economic journals. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table XII:
Omission and gender: peer effects

	Outcome variable: Omission				
	(1)	(2)	(3)	(4)	(5)
female j	0.021***	0.021***			0.017***
	(0.005)	(0.005)			(0.005)
Same Affiliation	-0.075***	-0.075***	-0.076***	-0.075***	
	(0.010)	(0.010)	(0.009)	(0.010)	
female j · Same Affiliation		-0.008			
		(0.052)			
At least one female j			0.010***	0.010***	
			(0.003)	(0.003)	
At least one female j · Same Affiliation				-0.003	
				(0.022)	
Connection					-0.098***
					(0.005)
female j · Connection					0.012
					(0.027)
N	88,175	88,175	102,664	102,664	88,175
R-sqr	0.066	0.066	0.068	0.068	0.071

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper emphasizing the effect of being in the same affiliation. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 5. *female j* represents paper j written by only women. The control variables as in the baseline. The variable *Same Affiliation* indicates if the citing paper and the cited or omitted paper have at least one of their authors that share the same affiliation. The variable *Connection* indicates if the citing paper and the cited or omitted paper have at least one of their authors that are either in the same affiliation, either coauthors, or coauthors of coauthors. The equations are estimated using a linear probability model. This regression excludes the self-citation. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Table XIII:
Omission and gender: androgynous versus female

	Outcome variable: Omission				
	Adrogynous	Adrogynous	Adrogynous	Adrogynous	"non typical white"
	(1)	(2)	(3)	(4)	(5)
	Proba <0.4	Proba <0.5	Proba <0.6	Proba <0.7	
Baseline controls	-0.319**	-0.270*	-0.252*	-0.478***	-0.166
	(0.152)	(0.145)	(0.138)	(0.124)	(0.103)
N	5,300	5,300	5,300	5,300	5,300
R-sqr	0.142	0.142	0.142	0.144	0.142

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper, emphasizing the effect of the name connotation. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The estimation focuses on cases with only females omitted, distinguishing between names known as females with a higher probability and names known as females with a relatively lower likelihood. For example, column (1) presents the results for female names with a probability lower than 0.4, putting names known as female with probability greater than 0.4 as a reference. Therefore, the coefficient has to be read relative to the latter category. The last column presents a similar analysis comparing typical non-white names to white names. For example, Asian names will be in the typical non-white sample and European and American names in the typical white sample. The control variables are the same as in the baseline and include the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect, cosine. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table XIV:
Omission and gender: seniority

	Outcome variable: Omission						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
female j	0.205***	0.177***	0.145***	0.204***	0.185***	0.154***	0.144***
	(0.043)	(0.043)	(0.043)	(0.043)	(0.044)	(0.044)	(0.055)
NBER_j	-0.181***					-0.167***	
	(0.026)					(0.026)	
max top5_j		-0.021***					
		(0.002)					
max papers_j			-0.014***			-0.014***	-0.013***
			(0.001)			(0.001)	(0.001)
superstar_j				-0.222***		-0.128***	
				(0.033)		(0.035)	
senior age_j					-0.008***	0.006***	
					(0.002)	(0.002)	
female $j \cdot \text{max papers}_j$							-0.000
							(0.010)
N	110,763	110,763	110,763	110,763	110,763	110,763	92,100
R-sqr	0.076	0.076	0.077	0.076	0.075	0.078	0.075

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper, emphasizing the effect of the seniority. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The other control variables are the same as in the baseline and include the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table XV:
Comparison with other fields: economics vs mathematics vs sociology

	(1) Economics	(2) Mathematics	(3) Sociology	(4) Comparison
female	0.137***	0.188***	0.113***	
	(0.023)	(0.035)	(0.024)	
mixed	-0.001	-0.037**	-0.017	
	(0.015)	(0.019)	(0.025)	
Math*Female				0.058
				(0.057)
Socio*Female				-0.11**
				(0.042)
j control	Y	Y	Y	Y
i controls	Y	Y	Y	Y
Time controls	Y	Y	Y	Y
Cross Field control				Y
N	750,509	324,568	278,495	596,844
R^2adj	0.044	0.101	0.089	0.098

This table shows the relationship between the omission and the gender of the authors in the cited or omitted paper, emphasizing the difference between Economics, Mathematics and Sociology. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The other control variables are the same as in the baseline and include the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by papers and reported in parentheses. To make the proper comparison, all cosine were computed using the fully unsupervised method. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table XVI:
History of omissions and future publications

	Outcome Variable			
	(1)	(2)	(3)	(4)
	Publication in Top 5	Citation (Asinh)	Innovativeness	"One shot"
History of Omission	-0.028***	-0.072***	-0.003***	0.035***
	(0.003)	(0.008)	(0.001)	(0.005)
Average Past citation	0.007*	0.167***	-0.001	-0.045
	(0.004)	(0.011)	(0.002)	(0.003)
Average Past Innovativeness index	0.030*	0.225***	0.027***	-0.048
	(0.015)	(0.050)	(0.008)	(0.038)
History of Top 5	0.034***	0.043***	0.003***	
	(0.003)	(0.004)	(0.001)	
Female	-0.003	-0.014	-0.000	0.066***
	(0.009)	(0.028)	(0.004)	(0.010)
Mean of dependent variable	0.330	2.168	0.671	0.532
N	25,529	25,529	25.529	15,673
R-sqr	0.243	0.422	0.680	0.156

This table presents the relationship between the history of past omissions and the probability to publish in a top 5 journal within the next 3 years. The history of omissions, $H_$Omissions, is measured as the inverse hyperbolic sine (asinh) of the cumulative number of past omissions. The control variables include author-specific variables, such as gender, characteristics of past publications (quality, citations, number of top 5), paper characteristics such as field, year of publication, time length (lag) between two publications. The regressions (1) and (4) reflect a linear probability model. For regression (4), "History of omission" characterizes the average number of omission of an author in the database. Regression (4) additionally controls for the seniority of the author in the profession as proxy by the current period minus the time of first publication. Standard errors clustered at author level and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Figures

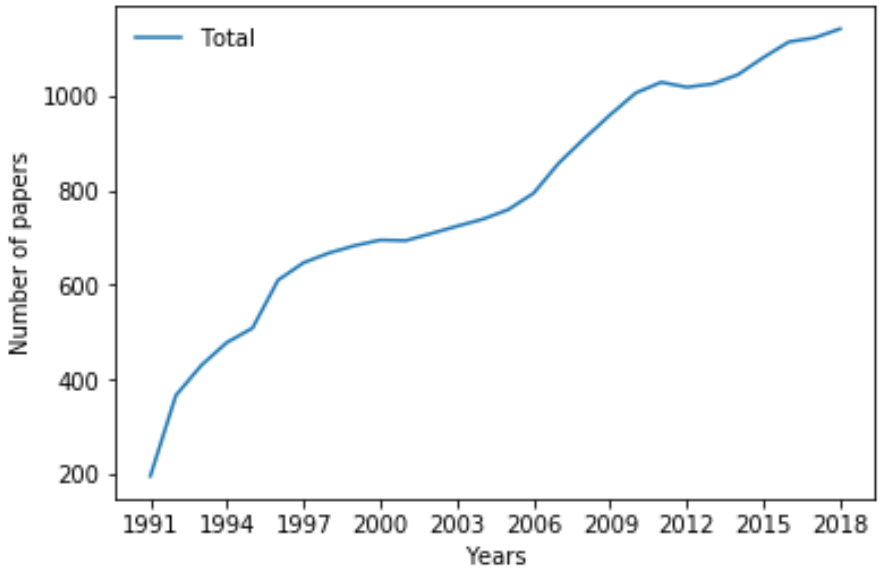


Figure I:
Number of articles over time

The figure shows the evolution of the number of articles over time from 1991 to 2018. The sample included publications in the 16 economic journals retained in this analysis: American Economic Review, Econometrica, Journal of Econometrics, Journal of economic growth, Journal of economic literature, Journal of economic perspectives, Journal of Economic Theory, Journal of Finance, Journal of Financial Economics, Journal of International Economics, Journal of Labour Economics, Journal of Monetary Economics, Journal of Political Economics, Quaterly Journal of Economics, Review of Economics Studies, Review of Financial Economics. The selected papers exclude proceedings papers, comments, articles of less than pages, books reviews, bibliographical items, articles without references and without abstracts, editorial material, letters and corrections.

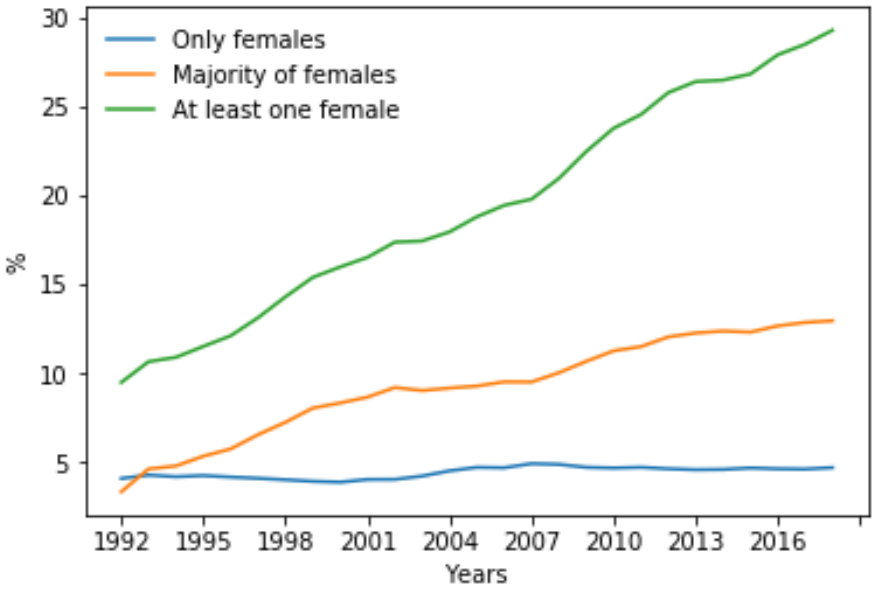


Figure II:
Female authors

The figure shows the evolution (five-year moving averages) of the share of articles with at least one female authors for a given year (green line), the share of articles with at least one female authors with a majority of female authors (orange line) and only female authors (blue line). The share of a certain category is the total number of articles in this category (at least one female author, majority of female author, only female authors) over the total number of articles.

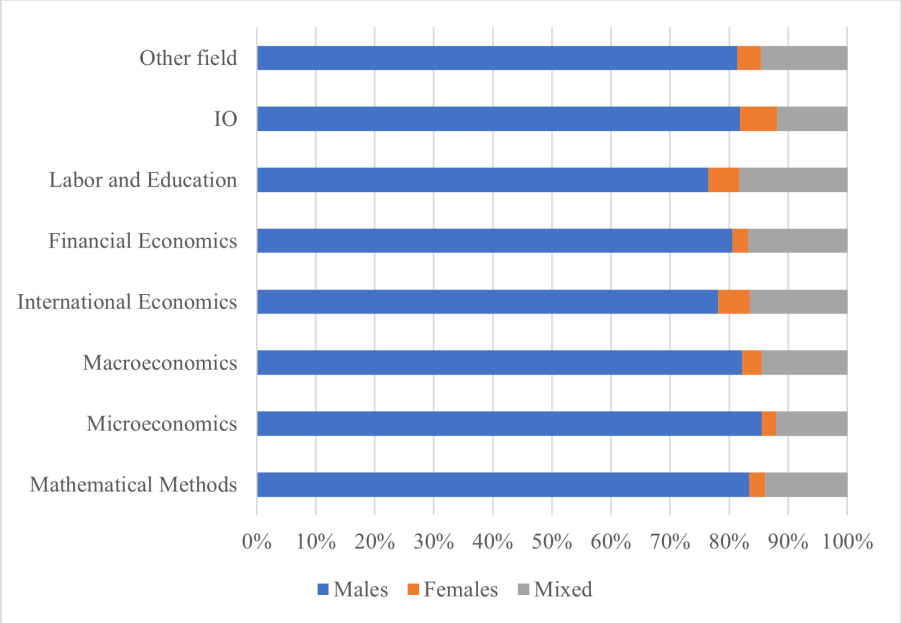


Figure III:
Field and references

This figures presents the share of references that are attributed to a given gender by field. Basically, the plot answers to the question: what is the average fraction of citations that refers to papers written by men, women or mixed gender? *Male* designed paper written by only men; *female* designed paper written by only women; *Mixed* designed paper written by a team of females and males.

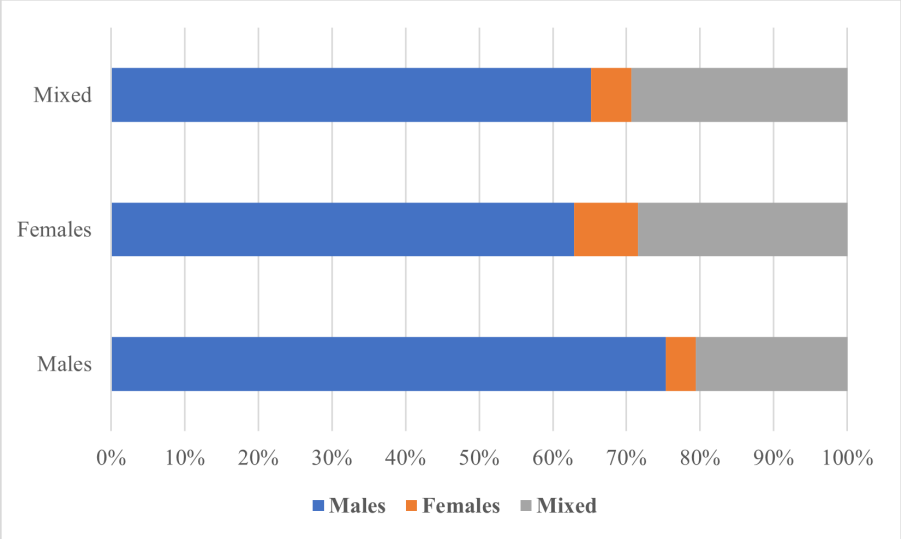


Figure IV:
Fraction of references to a given gender by gender of the citing paper

This figure presents the share of references that are attributed to a given gender depending on the "gender" of the citing paper. *Male* designed paper written by only men; *female* designed paper written by only women; *Mixed* designed paper written by a team of females and males.

Raw Data $\overset{\text{Data Cleaning}}{\Longrightarrow}$ **Tokens (Set of words)**

	paper p	paper p'	paper p''
trade_credit	6	7	5
retailer	5	1	1
supplier	6	4	1
purchase	3	3	0
bank	1	1	0
cash_constrain	1	1	0
creditor	1	1	0
finance	1	2	1
prediction	1	1	0
free_rider	3	0	0
competition	3	0	1
product_substitutability	2	0	0
...	...	...	...
Total	...	...	...

(a) Step 1

w: word ; p: paper, N: total number of papers in the database

TFIDF (w,p) $\times$

Term Frequency

$$\text{TF(w,p)} = \frac{\text{Number of } w \text{ in } p}{\text{Total Number of words in} p}$$

Inverse Document Frequency

$$\text{IDF(w)} = \frac{\text{Number of papers with } w}{N}$$

	paper p	paper p'	paper p''
trade_credit	$\frac{6}{Totalp}$* IDF(trade_credit)	$\frac{7}{Totalp}$* IDF(trade_credit)	$\frac{5}{Totalp}$* IDF(trade_credit)
retailer	$\frac{5}{Totalp}$* IDF(retailer)	$\frac{1}{Totalp}$* IDF(retailer)	$\frac{1}{Totalp}$* IDF(retailer)
...	...	...	...

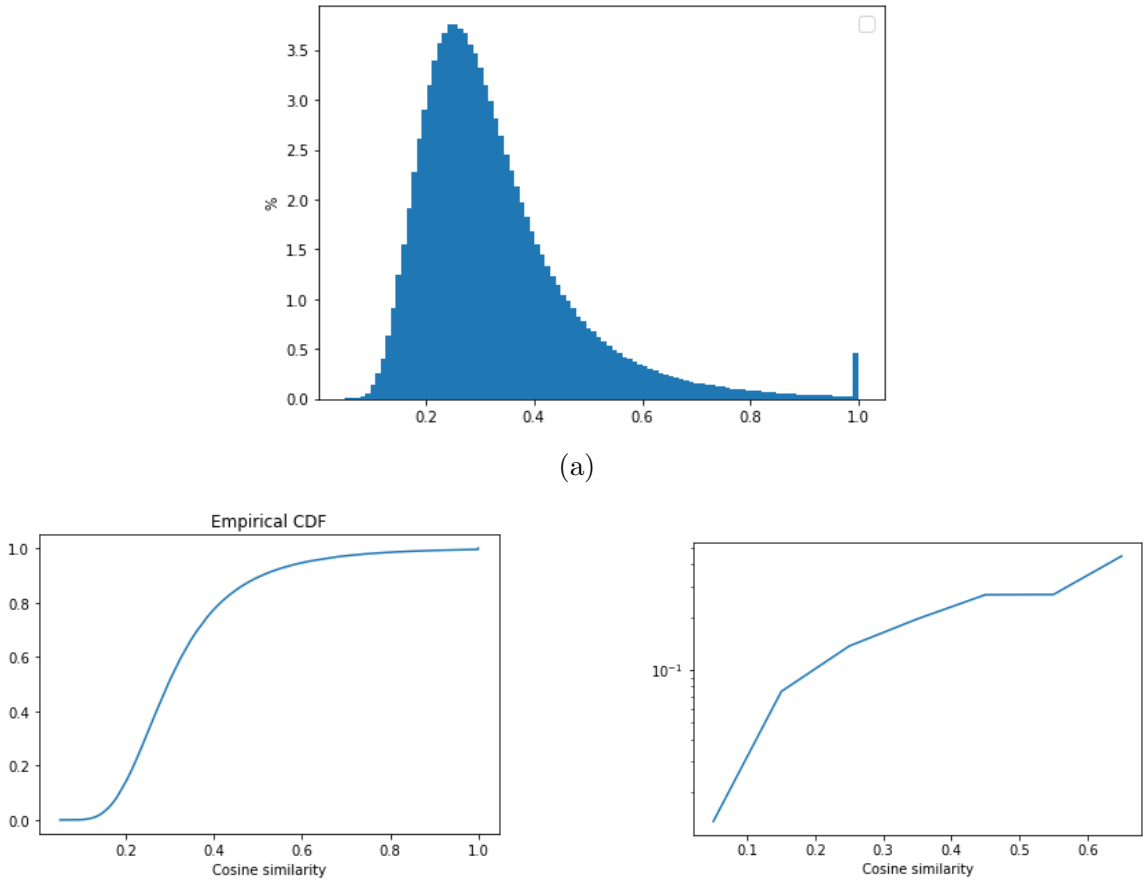
(b) Step 2

	p: Citing paper (2019)	p': Female paper **Omitted** (2010)	p'': Male paper **cited**(2004)
trade_credit	6	7	5
retailer	5	1	1
supplier	6	4	1
purchase	3	3	0
bank	1	1	0
cash_constrain	1	1	0
creditor	1	1	0
finance	1	2	1
prediction	1	1	0
free_rider	3	0	0
competition	3	0	1
product_substitutability	2	0	0
...	...	...	...
Total	...	...	...

(c) Step 3

Figure V:
Example

This figure illustrates the construction of the cosine similarity and the omission index. Panel (a) shows a part of the document-term matrix for a sample of three papers. Panel (b) presents the construction of the TF-IDF. Panel (c) presents the construction of the omission index.



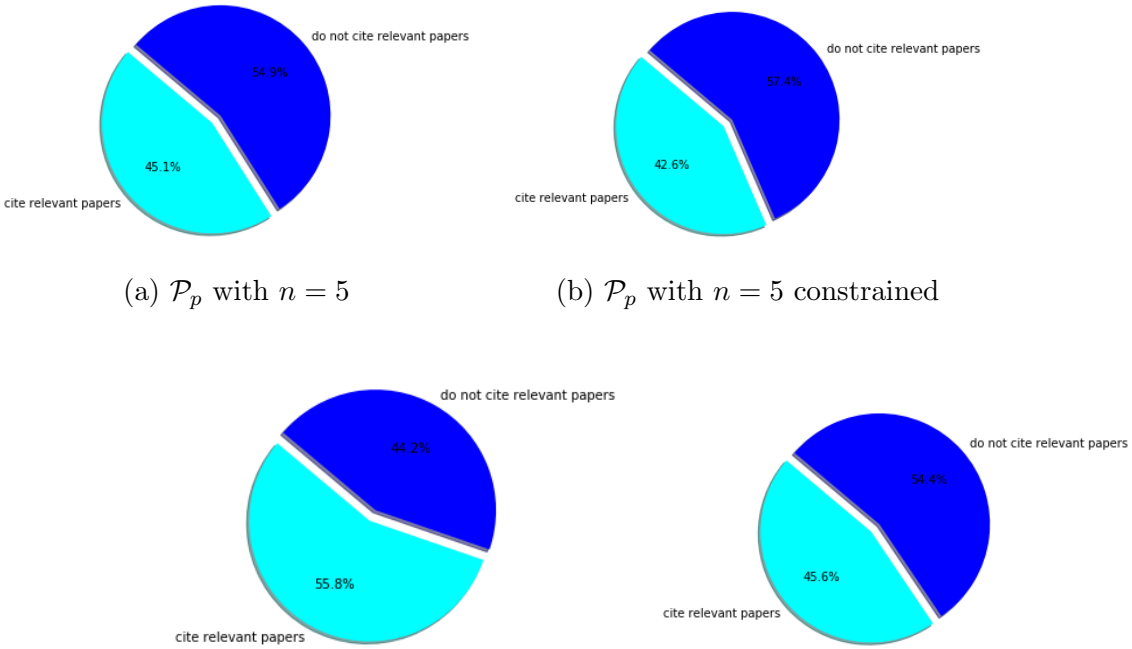
(a)

(b)

(c)

Figure VI:
Pairwise similarity

Panel (a) plots the distribution of the relative cosine. For paper P and P_{max} such as: $P_{max} = argmax_{P' \in C} cos(P, P')$, for any given article P' in the database, the relative cosine of paper P and paper P' is defined as:$\tilde{\lambda}_{p,p'} = \frac{\lambda_{p,p'}}{\lambda_{p,p_{max}}}$. Panel (b) shows the empirical cumulative distribution function (CDF) of the relative cosine. Panel (c) shows the conditional probability that paper p cites a paper p' as a function of the similarity score between those two papers. The computation excludes similarity score lower than 0.05.



(a) $\mathcal{P}_p$ with $n = 5$

(b) $\mathcal{P}_p$ with $n = 5$ constrained

(c) $\mathcal{P}_p$ with $n = 10$

(d) $\mathcal{P}_p$ with $n = 10$ constrained

Figure VII:
Aggregate overview of the omission propensity

The figure gives an aggregate overview of the propensity to omit relevant prior literature $\mathcal{P}_p$. A paper will be said to omit its relevant prior literature if it does not cite at least one paper from its relevant prior literature. At the opposite, a paper will be said to cite its relevant prior literature if it cites at least one paper from its relevant prior literature. Pie chart (a) shows the propensity of omission when $\mathcal{P}_p$ has 5 elements. Pie chart (b) shows the propensity of omission when $\mathcal{P}_p$ has 5 elements. For paper P and P_{max} such as: $P_{max} = argmax_{P' \in C} cos(P, P')$, for any given article P' in the database, the relative cosine of paper P and paper P' is defined as: $\tilde{\lambda}_{p,p'} = \frac{\lambda_{p,p'}}{\lambda_{p,p_{max}}}$. In the constrained specification, the relative cosine should be greater than 0.5. Similarly, Pie chart (c) shows the propensity of omission when $\mathcal{P}_p$ has 10 elements. Pie chart (d) shows the propensity of omission when $\mathcal{P}_p$ has 5 elements, but the relative cosine should be greater than 0.5.

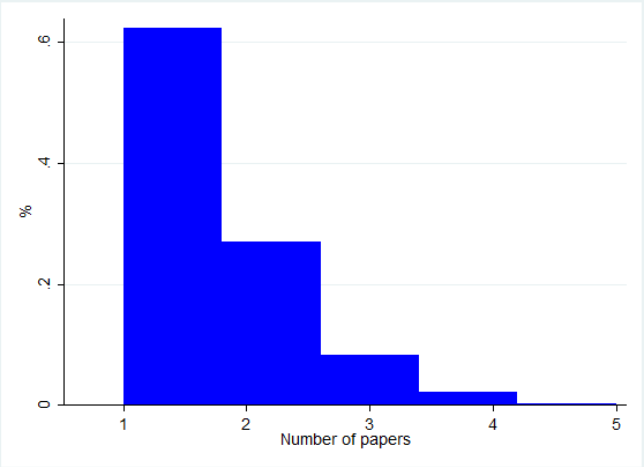
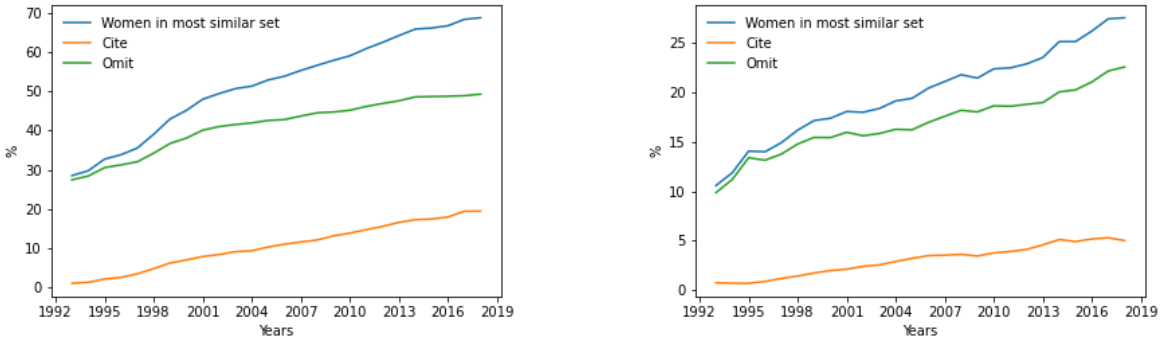


Figure VIII:
Number of paper cited among relevant prior literature

This figure shows the distribution of the number of papers in the relevant prior literature that are cited for the period 1991-2019. Relevant prior literature is determined by the five most similar papers to a given article as defined by the relative cosine.



(a) At least one female author

(b) Majority of female authors

Figure IX:
Prior literature with female authors

Panel (a) shows the fraction of papers with at least one female in their relevant prior literature (blue line), the fraction of papers that cites at least one paper with a female author (orange line), the fraction of papers that does not cite at least one paper with a female author (green line). Panel (b) shows the fraction of papers that have at least one paper with a majority of females in their relevant prior literature (blue line), the fraction of papers that cites at least one paper with a majority of females in their relevant prior literature (orange line), the fraction of papers that does not cite papers with a majority of females in their relevant prior literature (green line). The curves are five-year moving average and normalize by the 1993 values.

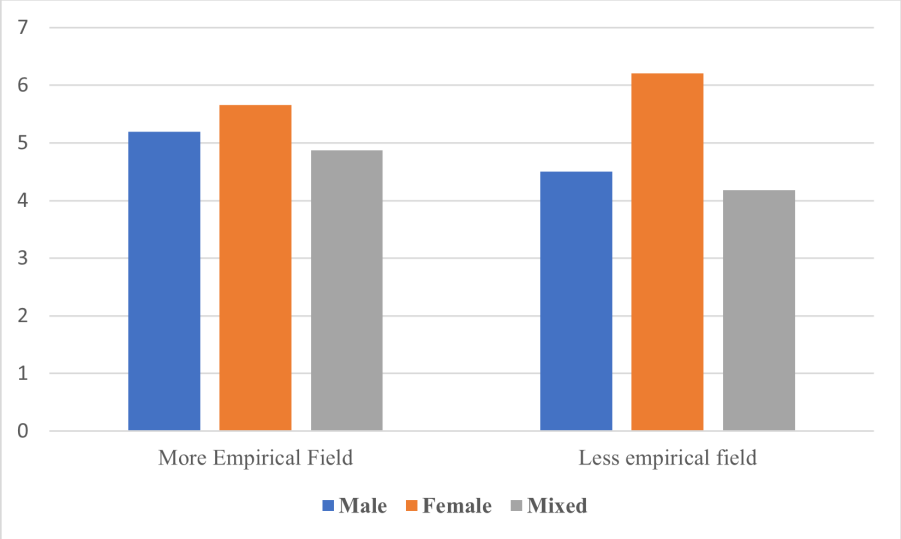
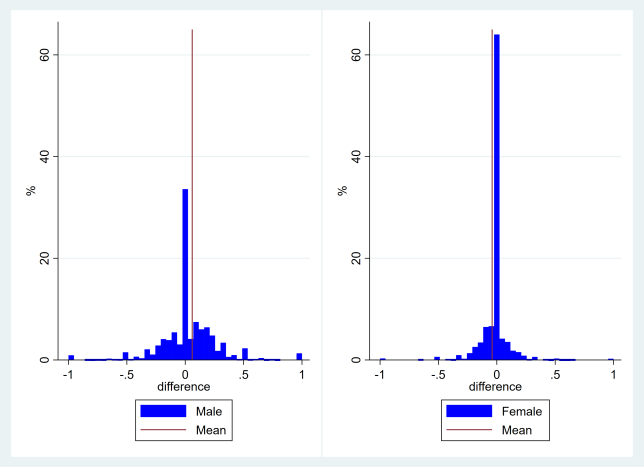
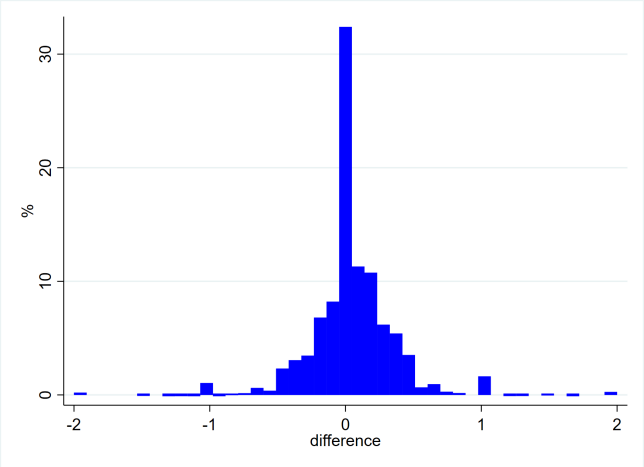


Figure X:
Odds of omission by field

Panel (a) decomposes the likelihood of being omitted compared to the one of being cited with respect to the gender and the field. *Male* designed paper written by only men; *female* designed paper written by only women; *Mixed* designed paper written by a team of females and males.



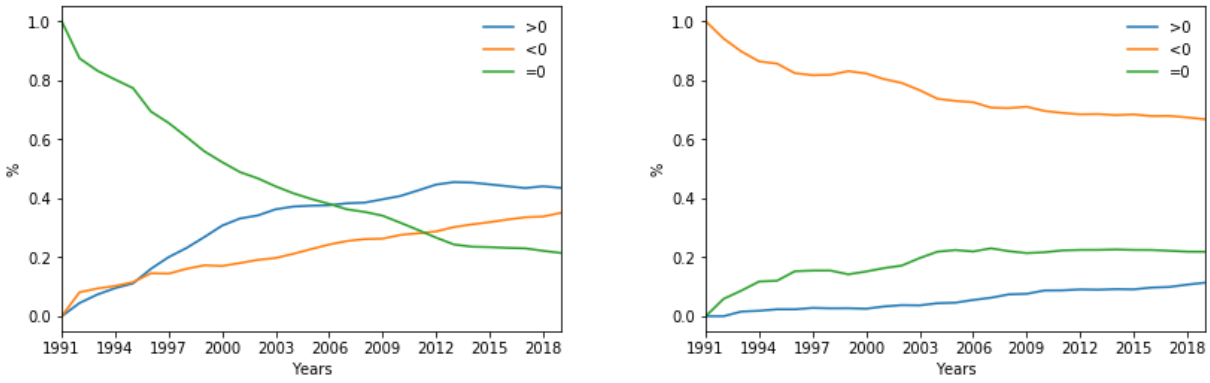
(a) Males versus Females



(b) Difference Males-Difference Females

Figure XI:
Actual distribution-target distribution (1/2)

This figure plots the difference between the actual distribution and the target distribution of gender type g. For each paper p, the actual distribution of gender is the share of papers in its references belonging to each category of gender (only males, only females, mixed gender). The target distribution of gender is the share of the closest papers (in the sense of the relative cosine) in its prior literature belonging to each category of gender (only males, only females, mixed gender). For each paper, the difference between the actual distribution of gender g and the target distribution is taken. Panel (a) plots the distribution of this difference for males versus females. A positive difference means that the actual distribution of a certain gender type is higher than the target distribution of this gender type. A negative difference means that the actual distribution of a certain gender type is lower than the target distribution of this gender type. Finally, a null difference means that the actual distribution of a certain gender type is the same as the target distribution of this gender type. Panel (b) takes the difference of the difference for males and for females. A positive double difference means that males are more *over-cited* or less *under-cited* than females.



(a) Males

(b) Females

Figure XII:
Actual distribution-target distribution (2/2)

This figure plots the difference between the actual distribution and the target distribution of gender type g. For each paper p, the actual distribution of gender is the share of papers in its references belonging to each category of gender (only males, only females, mixed gender). The target distribution of gender is the share of the closest papers (in the sense of the relative cosine) in its prior literature belonging to each category of gender (only males, only females, mixed gender). For each paper, the difference between the actual distribution of gender g and the target distribution is taken. Panel (a) and (b) plot the evolution of the share of papers with a positive, negative and null difference with respect to gender type male, female and mixed gender teams.

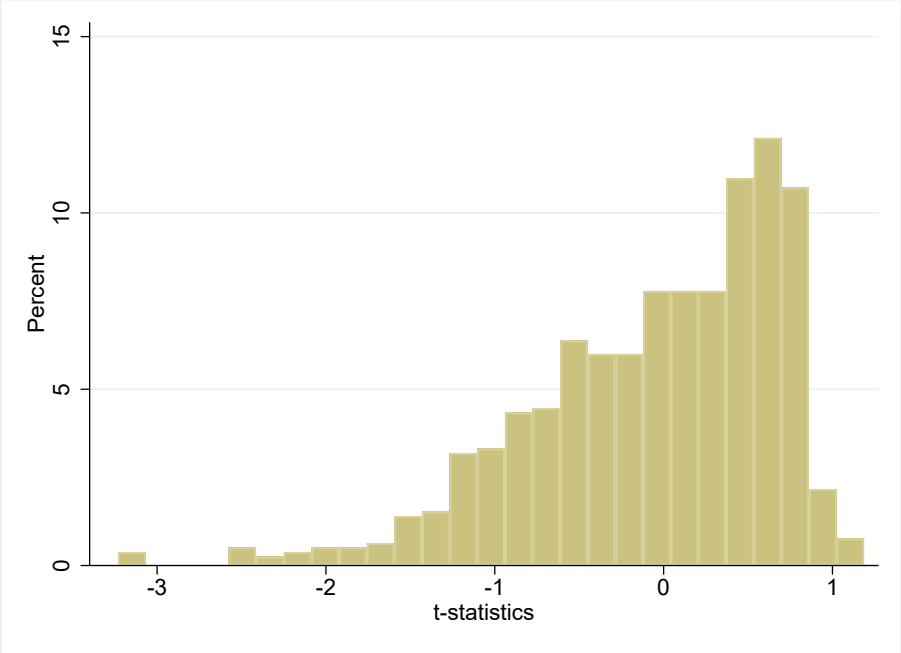
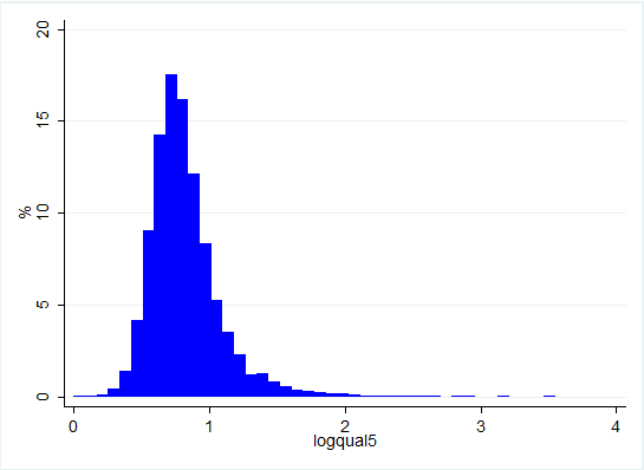
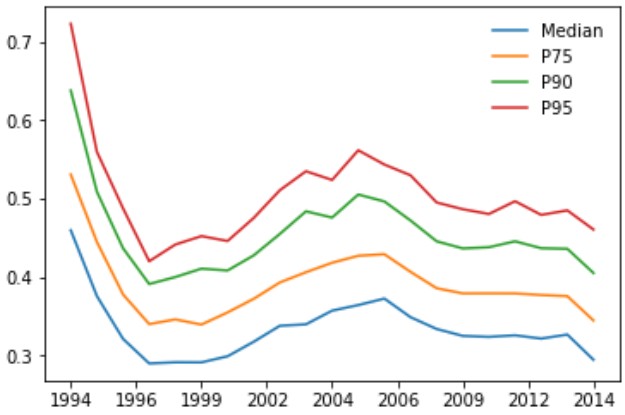


Figure XIII:
Placebo tests

The figure plots the distribution of the t-statistics associated with the all-female authored papers obtained after set of 1,000 random draws of the most similar set.



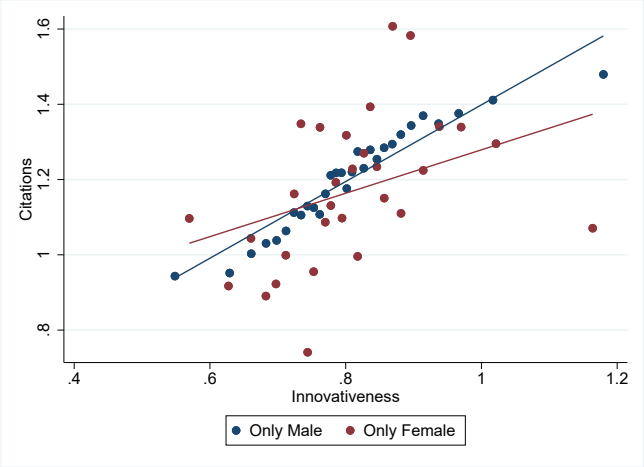
(a) Overall distribution



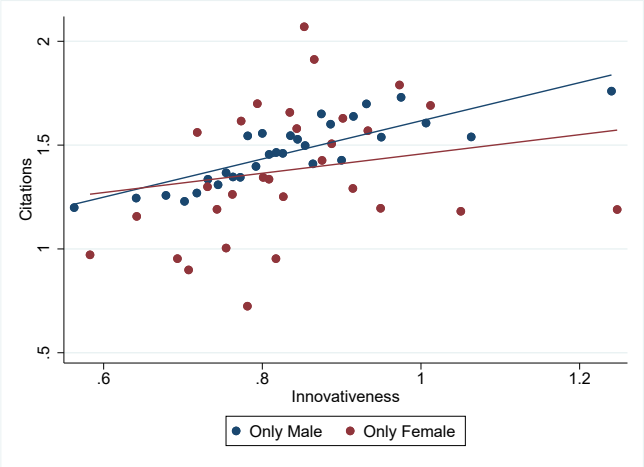
(b) Distribution over time

Figure XIV:
Distribution of innovativeness (quality) index

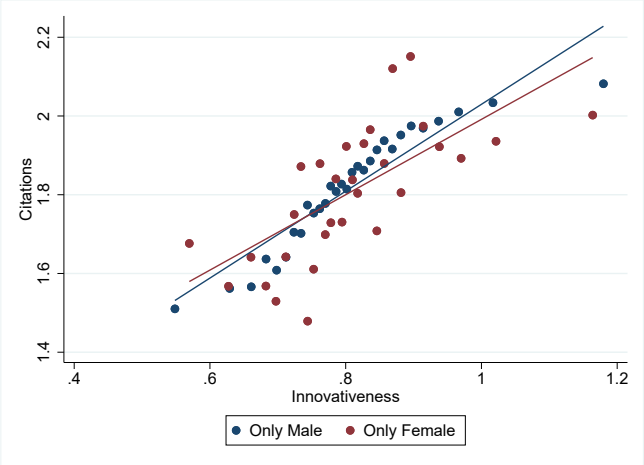
Panel (a) shows the overall distribution of the innovativeness index (q-index). Panel (b) shows the distribution of this index over time. The index is built following equation 6.2.2.



(a) Overall



(b) Only Top 5 Publications



(c) Counterfactual: citation compensating with omission cases

Figure XV:
Distribution of innovativeness (quality) index

The figure plots the link between the number of citations and the innovativeness index for papers written by males and females. The binned scatter plot controls for journals, field, institutions, year of publications, number of authors, maximum number of publications.

Appendix

A model of Citation choice

This section sets up a simple model to explain citation choice. The present model will underly the critical findings in the empirical part and help understand possible mechanisms in action. Therefore, it describes the context in which we could observe (1) a taste-based behavior, (2) a strategic decision, and (3) a statistical (or incorrect statistical) discrimination.

Preliminary: The model focuses on the citing paper and assume that there is no information issue in knowing the existence of a given paper. Authors want to maximize their chance to get published in a top journal. In doing so, they will maximize the quality of their article (related to the context, the topic, the methods) and the non-intrinsic factors (related to other components external to the quality per se). Let us focus on the references as a component of the non-intrinsic factors.

In maximizing this component, the author wants to choose a reference that satisfies either or all the three criteria: (1) a paper of quality, (2) a paper closely related, and (3) not missing a paper of a potential reviewer (referees, somebody in the editorial board,...).

Set up: Formally, let us consider two types of papers published before the citing paper i:a female paper f and a male paper m. For each i, we consider the set of papers I composed of $I = f, g$. From the citing paper i perspective, a given cited or omitted paper j is characterized by a perceived quality $E[Q]_{ij}$, an indicator of closeness d_{ij}, and how likely i think that j could be a reviewer Pr_{ij}. i may incur a cost in citing j that we note as c_{ij}.

Authors from i maximize the following payoff in choosing who to cite:

$$max_I \pi_{i,j} = \alpha E[Q]_{ij} + \beta d_{ij} + \gamma Pr_{ij} - c_{ij} + \nu_{ij} \tag{10}$$

We assume that $\alpha \geq 0$, $\beta \geq 0$, and $\gamma \geq 0$. ν_{ij} is an idiosyncratic factor like authors attach to other people in the field for example.

Under the assumption that ν_{ij} is independently and identically distributed according to the

extreme value distribution, The decision of who to cite is a logit model:

$$P(Cite_j^i = 1) = \frac{e^{(\alpha E[Q]_{ij} + \beta d_{ij} + \gamma Pr_{ij} - c_{ij})}}{\sum_{j' \in I} e^{(\alpha E[Q]_{ij'} + \beta d_{ij'} + \gamma Pr_{ij'} - c_{ij'})}} \tag{11}$$

Therefore, with $P_f = P(Cite_f^i = 1)$ the probability that i chooses to cite f and $P_m = P(Cite_m^i = 1)$ the probability that i chooses to cite m :

$$\frac{P_f}{P_m} = e^{(\alpha E[Q]_{if} + \beta d_{if} + \gamma Pr_{if} - c_{if}) - (\alpha E[Q]_{im} + \beta d_{im} + \gamma Pr_{im} - c_{im})} \tag{12}$$

Discussions:

1. Assume same quality and same relative distance: $E[Q]_{if} = E[Q]_{im}$, $d_{if} = d_{im}$
   Then, to have $P_f > P_m$, it has to be the case that $(\gamma Pr_{if} - \gamma Pr_{im}) - (c_{if} - c_{im}) > 0$.

   If $c_{if} = c_{im}$ (no difference in the cost of citing), then $P_f > P_m$, if $Pr_{if} > Pr_{im}$ meaning that i anticipate reaching one potential reviewer by citing f compared to citing m. In a subfield where the share of female is greatly lower than the share of male, this is unlikely. Therefore P_{im} will be greater than P_{if}.

   If $c_{if} > c_{im}$ (higher cost of citing female; case of high preference for males, taste-based discrimination against female or statistical discrimination against female): Then $P_f > P_m$, if $\gamma Pr_{if} - c_{if} > \gamma Pr_{im} - c_{if}$. In a subfield where females' share is considerably lower than the share of males, no women are cited in equilibrium. If the cost is associated with taste-based discrimination, a taste-based discrimination pattern could be sustainable in such a subfield. In a subfield where women are "at the top", we can assume without loss of generality that $Pr_{if} > Pr_{im}$.[25] Then, the results will depend on how one values the potential reviewer effect compared to the cost. If the reviewer effect is more elevated, any difference in costs,

[25]Because of the gender composition in Economics, there are almost no subfields where we can have $Pr_{if} > Pr_{im}$ in absolute. But without loss of generality, we can assume this is true in some subfields where the ratio Pr_{if}/Pr_{im} is relatively higher.

especially those related to taste-based discrimination, will not be sustainable in equilibrium.

If $c_{if} < c_{im}$, it means that there is a higher cost of citing males, or increased preference for females, or taste-based discrimination against male, or statistical discrimination against male. Thus in a subfield where females' share is much lower than the share of males, $P_f > P_m$ when the authors value less the reviewer effect or have a strong preference for citing females.

2. Assume same quality and same costs: $E[Q]_{if} = E[Q]_{im}$, $c_{if} = c_{im}$

   Let us consider four types of agents in this category: female closely related f^c, female not closely related f^l, male closely related m^c, and male not closely related m^l. For $P_{f^c} > P_{m^L}$, then $\beta d_{f^c} + \gamma Pr_{f^c} > \beta d_{m^l} + \gamma Pr_{m^l}$ (dropping i index for simplicity). We already have that $d_{f^c} > d_{m^l}$. Then, cases where $P_{f^c} < P_{m^L}$, thus m^L is cited and not f^c are sustainable in equilibrium as long as $Pr_{f^c} < Pr_{m^L}$, ie, one anticipates that because of the gender structure of the field, there is a greater likelihood for m to be a reviewer.

Data

Data cleaning

Several steps were used to process the textual information: remove the stopwords, drop punctuations, lower letter, lemmatize, tokenize. Stopwords are words commonly used such as "a","the"... A list of stopwords is built using information available online and internal to the software.[26] Lemmatazing or stemmatizing means put the words at their root. For example, "taxing" becomes "taxe". Finally, by tokenizing, the paragraphs are split into sentences and the sentences are split into words.

Institution Ranking

Information about author institutions were retrieved using the web of science database. When the data is available, each author could have one or many affiliations. All the affiliations available were extracted. Thereafter, the number of papers in the database that refer to a particular affiliation can be found. To avoid changes in university rankings, a more conservative method as in Engel (2019) was used. The institutions are ranked according to the number of papers they have in the database. When there are multiple coauthors, the paper's affiliation is taken to be the affiliation of the coauthor at the highest-ranked institution.

Primary Field

Articles are classified based on the Journal of Economic Literature Code (JEL) codes. As mentioned above the database of web of science does not include jel code. The one of Ideas Repec include the JEL codes of certain articles. Moreover, the articles usually have several JEL codes. The classification is done on the articles of the 16 newspapers selected over the period 1991-2019. Areas for the classification include: Microeconomics, macroeconomics, public finance, labor, industrial organization, development, urban economics, environmental, econometrics, finance, international, experimental (lab), economic history, economic economy, productivity,

[26] See for example $https://www.nltk.org/nltk_data/$.

law and economics, and other. To assign each article a primary, the following machine learning algorithm is considered. The dataset is split in two: a training dataset and a using dataset. With the training dataset, the algorithm can recognize the characteristics of the different categories. This dataset is composed of articles that have a single primary JEL code. For articles in newspapers whose the field is widely admitted (ex Journal of Labor economics), the field of the journal is assigned (see Angrist et al. (2019)). To avoid a high level of success due to an over-representation of a class, the sample is chosen to ensure proportional share of each field. The idea is to predict the field using the characteristics of the article. The dependent variables will therefore be the titles and the keywords.[27]

The training database was used to train a couple of classifier including the naive Bayes, the logistic, the random forest classifier, and the Support Vector Classifier (SVC) using the titles and keywords to predict the JEL code. The package used is the "Scikit-learn" package (Pedregosa et al., 2011).[28] Titles and keywords were subject to a cleaning procedure (remove stopwords, consider ngrams, drop punctuations, lemmatize ...) and transform into digital data by the TF-IDF. We used a grid search to get the optimal hyper-parameter values for the classification. Thus, in a subset with 90-10 training-test sample, the accuracy of the algorithm is estimated to approximately 90% using the SVC.

[27]One can also use the abstracts. The results remain the same.

[28]The SVC classifier was the best classifier among a set of tests made with other classifiers like Random Forest, decision trees ...

Example of similar papers identified by the algorithm

Cited paper	Citing paper
Auer, R. A., & Schoenle, R. S. (2016). Market structure and exchange rate pass-through. *Journal of International Economics*, *98*, 60-77.	Fitzgerald, D., & Haller, S. (2014). Pricing-to-market: evidence from plant-level prices. *Review of Economic Studies*, *81*(2), 761-786.
Basu, A. K., & Chau, N. H. (2004). Exploitation of child labor and the dynamics of debt bondage. *Journal of Economic Growth*, *9*(2), 209-238.	Baland, J. M., & Robinson, J. A. (2000). Is child labor inefficient?. *journal of Political Economy*, *108*(4), 663-679.
Beltratti, A., & Morana, C. (2006). Breaks and persistency: macroeconomic causes of stock market volatility. *Journal of econometrics*, *131*(1-2), 151-177.	Campbell, J. Y., Lettau, M., Malkiel, B. G., & Xu, Y. (2001). Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *The journal of finance*, *56*(1), 1-43.
Bessembinder, H., & Seguin, P. J. (1992). Futures-trading activity and stock price volatility. *the Journal of Finance*, *47*(5), 2015-2034.	Daigler, R. T., & Wiley, M. K. (1999). The impact of trader type on the futures volatility-volume relation. *The Journal of Finance*, *54*(6), 2297-2316.
Bombardini, M., Orefice, G., & Tito, M. D. (2019). Does exporting improve matching? Evidence from French employer-employee data. *Journal of International Economics*, *117*, 229-241.	Krishna, P., Poole, J. P., & Senses, M. Z. (2014). Wage effects of trade reform with endogenous worker mobility. *Journal of International Economics*, *93*(2), 239-252.
Borjas, G. J., & Hilton, L. (1996). Immigration and the welfare state: Immigrant participation in means-tested entitlement programs. *The quarterly journal of economics*, *111*(2), 575-604.	Bratsberg, B., Raaum, O., & Røed, K. (2010). When minority labor migrants meet the welfare state. *Journal of Labor Economics*, *28*(3), 633-676.
Brown, J., Duggan, M., Kuziemko, I., & Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program. *American Economic Review*, *104*(10), 3335-64.	Aizawa, N., & Kim, Y. S. (2018). Advertising and risk selection in health insurance markets. *American Economic Review*, *108*(3), 828-67.
Callaghan, S. R., Saly, P. J., & Subramaniam, C. (2004). The timing of option repricing. *The Journal of Finance*, *59*(4), 1651-1676.	Carter, M. E., & Lynch, L. J. (2001). An examination of executive stock option repricing. *Journal of Financial Economics*, *61*(2), 207-225.
Chernozhukov, V., & Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, *132*(2), 491-525.	Horowitz, J. L., & Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica*, *75*(4), 1191-1208.
Choi, Y. S., & Krishna, P. (2004). The factor content of bilateral trade: An empirical test. *Journal of Political Economy*, *112*(4), 887-914.	Lai, H., & Zhu, S. C. (2007). Technology, endowments, and the factor content of bilateral trade. *Journal of International Economics*, *71*(2), 389-409.
Damm, A. P. (2009). Ethnic enclaves and immigrant labor market outcomes: Quasi-experimental evidence. *Journal of Labor Economics*, *27*(2), 281-314.	Edin, P. A., Fredriksson, P., & Åslund, O. (2003). Ethnic enclaves and the economic success of immigrants—Evidence from a natural experiment. *The quarterly journal of economics*, *118*(1), 329-357.

Cited paper	Citing paper
Dubois, P., & Lasio, L. (2018). Identifying industry margins with price constraints: Structural estimation on pharmaceuticals. *American Economic Review*, *108*(12), 3685-3724.	Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica*, *69*(2), 307-342.
Dupor, B., & Guerrero, R. (2017). Local and aggregate fiscal policy multipliers. *Journal of Monetary Economics*, *92*, 16-30.	Barro, R. J., & Redlick, C. J. (2011). Macroeconomic effects from government purchases and taxes. *The Quarterly Journal of Economics*, *126*(1), 51-102.
Edmonds, E. V., & Pavcnik, N. (2005). Child labor in the global economy. *Journal of Economic Perspectives*, *19*(1), 199-220.	Baland, J. M., & Robinson, J. A. (2000). Is child labor inefficient?. *journal of Political Economy*, *108*(4), 663-679.
Geronimus, A. T., & Korenman, S. (1992). The socioeconomic consequences of teen childbearing reconsidered. *The Quarterly Journal of Economics*, *107*(4), 1187-1214.	Hotz, V. J., Mullin, C. H., & Sanders, S. G. (1997). Bounding causal effects using data from a contaminated natural experiment: Analysing the effects of teenage childbearing. *The Review of Economic Studies*, *64*(4), 575-603.
Harrigan, J., Ma, X., & Shlychkov, V. (2015). Export prices of US firms. *Journal of International Economics*, *97*(1), 100-111.	Manova, K., & Zhang, Z. (2012). Export prices across firms and destinations. *The Quarterly Journal of Economics*, *127*(1), 379-436.
Hatfield, J. W., & Milgrom, P. R. (2005). Matching with contracts. *American Economic Review*, *95*(4), 913-935.	Kojima, F., & Pathak, P. A. (2009). Incentives and stability in large two-sided matching markets. *American Economic Review*, *99*(3), 608-27.
Kang, J. K., & Shivdasani, A. (1995). Firm performance, corporate governance, and top executive turnover in Japan. *Journal of financial economics*, *38*(1), 29-58.	Denis, D. J., Denis, D. K., & Sarin, A. (1997). Ownership structure and top executive turnover. *Journal of financial economics*, *45*(2), 193-221.
Keane, M., & Stavrunova, O. (2016). Adverse selection, moral hazard and the demand for Medigap insurance. *Journal of Econometrics*, *190*(1), 62-78.	Fang, H., Keane, M. P., & Silverman, D. (2008). Sources of advantageous selection: Evidence from the Medigap insurance market. *Journal of political Economy*, *116*(2), 303-350.
Klette, T. J., & Kortum, S. (2004). Innovating firms and aggregate innovation. *Journal of political economy*, *112*(5), 986-1018.	Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. *The American economic review*, 562-583.
Knoeber, C. R., & Thurman, W. N. (1994). Testing the theory of tournaments: An empirical analysis of broiler production. *Journal of labor economics*, *12*(2), 155-179.	Levy, A., & Vukina, T. (2004). The league composition effect in tournaments with heterogeneous players: An empirical analysis of broiler contracts. *Journal of labor economics*, *22*(2), 353-377.
Levitt, S. D. (2004). Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. *Journal of Economic perspectives*, *18*(1), 163-190.	Engelhardt, B. (2010). The effect of employment frictions on crime. *Journal of Labor Economics*, *28*(3), 677-718.

Cited paper	Citing paper
Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. *Econometrica: Journal of the Econometric Society*, 711-730.	Lewbel, A., & Pendakur, K. (2009). Tricks with Hicks: The EASI demand system. *American Economic Review*, *99*(3), 827-63.
Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, *24*(2), 129-44.	Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter?. *Journal of Labor Economics*, *31*(3), 443-499.
Petersen, M. A., & Rajan, R. G. (1994). The benefits of lending relationships: Evidence from small business data. *The journal of finance*, *49*(1), 3-37.	Brown, J. D., & Earle, J. S. (2017). Finance and growth at the firm level: evidence from SBA loans. *The Journal of Finance*, *72*(3), 1039-1080.
Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.	Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of labor Economics*, *25*(1), 95-135.
Romano, J. P., & Wolf, M. (2001). Subsampling intervals in autoregressive models with linear time trend. *Econometrica*, *69*(5), 1283-1314.	Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, *75*(5), 1411-1452.
Servaes, H. (1996). The value of diversification during the conglomerate merger wave. *The Journal of Finance*, *51*(4), 1201-1225.	Lins, K., & Servaes, H. (1999). International evidence on the value of corporate diversification. *The Journal of Finance*, *54*(6), 2215-2239.
Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, *99*(1), 179-215.	Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*(4), 1239-1285.
Van Zandt, T. (1999). Real-time decentralized information processing as a model of organizations with boundedly rational agents. *The Review of Economic Studies*, *66*(3), 633-658.	Vayanos, D. (2003). The decentralization of information processing in the presence of interactions. *The Review of Economic Studies*, *70*(3), 667-695.
Whang, Y. J., & Andrews, D. W. (1993). Tests of specification for parametric and semiparametric models. *Journal of Econometrics*, *57*(1-3), 277-318.	Hong, Y., & White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica: Journal of the Econometric Society*, 1133-1159.
Zlate, A. (2016). Offshore production and business cycle dynamics with heterogeneous firms. *Journal of International Economics*, *100*, 34-49.	Bergin, P. R., Feenstra, R. C., & Hanson, G. H. (2011). Volatility due to offshoring: Theory and evidence. *Journal of International Economics*, *85*(2), 163-173.