ORIGINAL ARTICLE

Gender distribution across topics in the top five economics journals: a machine learning approach

J. Ignacio Conde-Ruiz[1,3] · Juan-José Ganuza[2] · Manu García[4] · Luis A. Puch[3]

© The Author(s) 2021

Abstract

We analyze text data in all the articles published in the top five (T5) economics journals between 2002 and 2019 in order to find gender differences in their research approach. We implement an unsupervised machine learning algorithm: the structural topic model (STM), so as to incorporate gender document-level meta-data into a probabilistic text model. This algorithm characterizes jointly the set of latent topics that best fits our data (the set of abstracts) and how the documents/abstracts are allocated to each topic. Latent topics are mixtures over words where each word has a probability of belonging to a topic after controlling by journal name and publication year (the meta-data). Thus, the topics may capture research fields but also other more subtle characteristics related to the way in which the articles are written. We find that females are unevenly distributed over the estimated latent topics. This and other findings rely on "automatically" generated built-in data given the contents in the abstracts of the articles in the T5 journals, without any arbitrary allocation of texts to particular categories (as JEL codes, or research areas).

✉ Juan-José Ganuza
  juanjo.ganuza@gmail.com

[1] Fedea, Madrid, Spain

[2] Universitat Pompeu Fabra and Barcelona GSE, Barcelona, Spain

[3] Universidad Complutense de Madrid and ICAE, Madrid, Spain

[4] Washington University in St. Louis and ICAE, St. Louis, MO, USA

Ⓐ Springer

1 Introduction

Despite the efforts undertaken for the whole economic profession to fight against discrimination, women are underrepresented in academia. Lundberg and Stearns (2019) make an assessment of the presence of female economists in the profession, and they report a very slow improvement in the last two decades. The picture is as follows. In the beginning of this century, 35% percent of PhD students and 30% of assistant professors were female. Since then, these numbers have not increased.[1] Additionally, Siniscalchi and Veronesi (2020) summarizing Chevalier (2020) (Report of the Committee on the Status of Women in the Economics Profession) point out that the proportion of women assistant professors in the "top 10" schools has declined to less than 20% by 2019. They document also that female have been less successful in promoting to tenured associate or full professors.

In economics, the tenure path often requires to publish in the top five (Top 5, or just T5) journals, namely *American Economic Review* (*AER*), *Econometrica* (*ECA*), *Journal of Political Economy* (*JPE*), *Quarterly Journal of Economics* (*QJE*) and *Review of Economic Studies* (*REStud*). Heckman and Moktan (2020) analyze the tenure decisions of the top 35 Economics departments in the USA, and they conclude that T5 publications are a very powerful explanatory variable of the promotion to tenure. Publishing in a T5 is becoming the main goal of young professors in economics because their professional career may depend on succeeding on this target. In addition, the content published in these journals is also determining the path of research in economics. As a consequence of these facts, the competition to publish in any of these journals has increased in recent years. Card and DellaVigna (2013) analyze the publication records in the Top 5 from 1970 to 2012 showing that the acceptance rate has fallen from 15% (1970) to 6% (2012). They explain this fact as a combination of the increasing number of submissions and a declining number of published papers. Card et al. (2019) further analyze the publication records from two of the T5 journals (the *QJE* and *REStud*), together with the *Journal of European Economic Association* and the *Review of Economics and Statistics*. They report that the current proportion of accepted papers is 3%. Is the T5 entry barrier harder for women? The answer provided by Card et al. (2019) to this question is ambiguous. On the one hand, these authors do not find any gender biases in the refereeing process, and editors decisions are gender-neutral conditional on the referee advises. On the other hand, they find that conditional on referee process,

---

[1] Boustan and Langan (2019) analyze the performance of women across PhD programs in economics. They report that in 2017, women were a 32% of entering PhD students in economics, This proportion of women in economics is below many other fields including science, technology, engineering, and mathematics (see also Bayer and Rouse (2016)).

female authored papers end up accumulating more citations in later years.[2] A potential explanation for this second result is that journals hold female-authored papers to higher standards. Hengel (2020) uses readability scores and finds that female-authored papers are better written and improve during peer review and as they publish more papers. These results could be related to some "horizontal" features or characteristics of female-authored papers that lead to more citations or better writing standards, but not to higher acceptance rates in the editorial process. As Card et al. (2019) control by research fields (JEL codes), their results may be linked to more subtle horizontal differences. For instance, in the same research field, males may choose a more theoretical approach and females a more applied perspective (which tends to be more cited or subject to less complicated wording), leading to particular career outcomes.[3]

Several papers have pointed out persistent gender differences in the choice of research fields in economics. Dolado et al. (2012) analyze the gender distribution of research fields in the top-50 economics departments in 2005, and show that women are unevenly distributed across fields. Similarly, Chari and Goldsmith-Pinkham (2017) use data from submissions to the National Bureau of Economic Research Summer Institute (2001–2016) and show that the distribution of female researchers is not uniform across fields. From these, we learnt that women are particularly underrepresented in macro, finance and economic theory, and more prevalent in labor or applied microeconomics fields. Beneito et al. (2021) find similar results using data from the annual AEA meetings from 2010–2016, while Lundberg and Stearns (2019) focus on PhD dissertations in Economics from 1991–2017, in almost all major PhD-granting departments in the USA. Using the JEL code for research areas, they find that women are more prone to study topics in Labor and Public Economics than in Macro and Finance. They also show that this pattern has not changed over time.

We want to contribute to this literature in two directions. First, we focus on exploring the gender horizontal distribution across research topics in the leading economics journals. We do so by using a new methodological approach based on machine learning techniques. This classifies our abstracts' database into *latent topics*. We collect all the articles published in T5 journals for the period 2002–2019. We obtain 5311 articles, and we keep track for each article of the authors' names, year of publication, journal and the abstract. With this information, we can provide a very accurate picture of the performance of men and women publishing record in these leading journals. Our primary objective is to describe what these latent topics are and the gender distribution across them. Notice this is a very particular sample of researchers though.

Second, from the universe of algorithms for topic modeling we implement and develop the structural topic model (STM) developed by Roberts et al. (2019). This

---

[2] Hengel and Moon (2020) analyze publications in T5 and they also find that female authors published articles are more cited.

[3] We borrow from the industrial organization literature the term "horizontal" differentiation since we refer to differences in gender approaches and topics choices unrelated to research quality ("vertical" differentiation) When those "horizontal" differences exist, papers of the same subjective quality may receive different citations depending for instance on the popularity of the topic or the number of scholars working on it. If we were able to control for "horizontal" gender differences (which is the goal of the paper), we could identify, in a more accurate way, the potential gender discrimination biases. We will discuss in greater detail, how to use our methodology for assessing gender discrimination biases when we discuss our future research agenda at the conclusion section.

choice is because the algorithm allows to incorporate document-level meta-data into a probabilistic text model. Precisely, we keep track of journal names and publication years as covariates to improve the estimation of the prevalence of topics in our data. Our abstracts come from different sources and different periods of time, so it is natural to allow this meta-data to affect the frequency with which a topic appears. The output of the algorithm is a stochastic model that generates latent topics and allocate the documents to them in a probabilistic way. The main advantage of this unsupervised machine learning approach is that latent topics are mixtures over words where each word has a probability to belong to the different topics. Therefore, these topics can capture, conditional on covariates and without human intervention, research fields, information regarding the style of writing, methodology, conversational patterns or even different ways of thinking.

We start by identifying the number of latent topics for which the stochastic model fits best our data. The result is that female authors are unevenly distributed across latent topics. It turns out that female prevalence dispersion is higher across these topics compared to other approaches. Moreover, we show that although the proportion of females is slightly increasing among the population of T5 authors over the years, the identified horizontal differences persist. We compute the empirical distribution of latent topics by gender and we show some striking differences between male and female expected proportions. We want to emphasize the importance of these results, not only because latent topics may capture subtle horizontal differences, but also because the gender differences we estimate are "automatically" generated given the documents, without any arbitrary allocations to particular categories (as JEL codes, or declared areas). Thus, they are possibly more robust.

Notwithstanding, the choice of the number of latent topics, even if optimal as we discuss, is subject to clustering issues. To address these issues, we also choose to reduce the number of topics the algorithm has to generate, and in order to capture the mixtures of words that more closely resemble to research areas. There is a trade-off when choosing ex ante the number of latent topics. On the one hand, a relatively high number of topics usually fits better the data. On the other hand, a lower number of latent topics facilitates the broad semantic interpretation of them. In our setting, a lower number of topics turns out to make them closer to traditional research fields. Consistently with our main findings, we corroborate the uneven distribution of topic/research fields by gender, but now, much more in line with the existing literature cited above. Thus, we can also discuss the link between the existing literature and our class of probabilistic results. Our approach provides complementary evidence from previous literature over horizontal research differences between males and females. The idea is that the larger set of research topics may allow to identify more precisely the gender gaps, and what is more important, may help to understand the driving forces behind these gaps.

There are several channels for which the gender differences in the choice of research topic that we identify can have an impact on the probability of publishing in top journals, earning tenure and in general on career success. Conde-Ruiz et al. (2017, 2021) and Siniscalchi and Veronesi (2020) provide two dynamic mechanisms that may explain how "horizontal" gender differences, together with an initially uneven distribution of gender researchers, may generate an *unintentional discrimination trap* linked with the functioning of academic organizations (journals, departments, etc.).

In particular, Conde-Ruiz et al. (2017, 2021) analyze a promotion setting in which workers' skills are assessed by committees whose members have different abilities to evaluate workers' signals (they are better at evaluating workers from the same group). This "homo-accuracy" assumption naturally translates to the present academic setting, where promotions and editorial processes are done by "committees" and where evaluators making research in the same research field are able to assess better the underlying quality of the candidate. Under this "*homo-accuracy bias*," the group that is most represented in the evaluation committee generates more accurate signals, and, consequently, has a greater incentive to invest in human capital. This gives rise to a discrimination trap. If, for some exogenous reason, one group is initially poorly evaluated (less represented into evaluation committees), this translates into lower investment in human capital of individuals of such group, which leads to lower representation in the evaluation committee in the future, generating a persistent discrimination process. Siniscalchi and Veronesi (2020) focus specifically on the academic labor market and point out a similar unintentional discrimination trap linked to the so-called *self-image bias*. Research evaluation is biased toward young researchers with similar characteristics to them. The authors build up an overlapping-generations model with two groups of researchers with equally desirable (but a little bit different) research characteristics and identical ex ante productivity distributions. If one group is slightly over-represented into the evaluation group, this group (and its specific research characteristics) may dominate forever. These theoretical results go in line with the empirical findings of Dolado et al. (2012) that show that the probability for a female researcher to work on a given field is positively related to the share of women already working on that field (path-dependence). The proportions these authors find based on JEL codes are very similar to what we find automatically at the same level of aggregation, but we can set forth a lot more field idiosyncrasy under an extended optimal topic choice. At the end of the paper, we discuss various issues for further research in related applications.

The paper is organized as follows: the next section presents the raw data and the descriptive analysis of the patterns of publication in T5 journals. Section 3 presents the structural topic model. Section 4 studies the gender differences in the latent estimated topics. Section 5 extends the model to analyze topics as research fields. Last section concludes, and in Appendix we explore several extensions and provide details about the functioning of the structural topic model (STM) algorithm.

2 Raw Data and Descriptive Analysis

We collect the publicly available information from all articles published between 2002 and 2019 in the T5 leading journals in economics, as already indicated: *The American Economic Review*, *Econometrica*, *The Journal of Political Economy*, *The Quarterly Journal of Economics*, and *The Review of Economic Studies*. For each article, we collect the information about the journal, year of publication, authors and the abstract of the paper.

We have 5311 articles in total over the period 2002–2019, the average number of papers published in top-5 journals per year is 295, with a maximum of 351 (on year 2017), and a minimum of 234 (on year 2002). Figure 1 shows that the distribution of
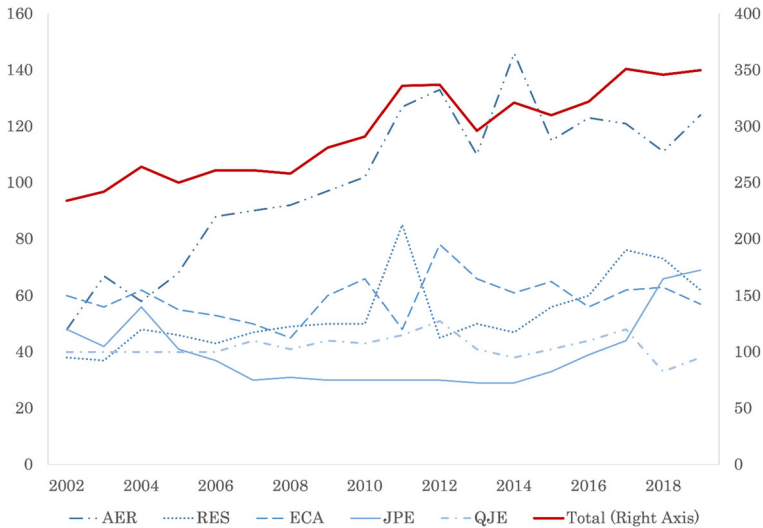
Fig. 1 Number of articles published per year in T5. *Note* Publications exclude notes (without abstract), comments, announcements, and Papers and Proceedings (P&P)

published papers by journal is uneven. *AER* accounts for 34.3%, while *JPE* only represent 13.4% of the sample. *AER* publishes regular articles as well as shorter papers.[4] We include in our sample the shorter papers (as long as they have abstract) since their editorial processes is similar to regular articles. We exclude the articles published in *AER* as Papers and Proceedings since their requirements and editorial processes are different.[5] We want to compare this descriptive information with Card and DellaVigna (2013) who analyze all the articles published in the T5 from 1970 to 2012. They obtain several interesting facts, among them, that the total number of articles published in these journals declined from 400 per year in the late 1970s to 300 per year in 2012. They also show that one journal, the *American Economic Review*, accounted in 2012 for 40% of T5 publications, up from 25% in the 1970s. In our updated sample, as it is shown in the figure, we find that this trend has stabilized after 2012.

Card and DellaVigna (2013) also find that the number of authors per paper has increased from 1.3 in 1970 to 2.3 in 2012. We observe the same trend in the recent years, in particular in 2019 the average number of authors was above 2.5. Figure 2 reports the share of articles by number of authors, one to five or more. Clearly, the steepest trend downward is for solo authorship, whereas the three-author case (or even the four-author case) exhibits the opposite pattern. The two-author case share has remained fairly stable over the entire sample at around 40% of articles (base, not augmented). Five or more authors in economics articles at leading journals are still a rare event.

Next, we move to analyze gender issues. We do not observe directly gender in our data. For solving that problem, we classify authors by gender according to their first

---

[4] *AER* stopped publishing shorter papers in 2018.

[5] In "Appendix E", we add P&P articles to our data and we replicate the analysis for these extended data.
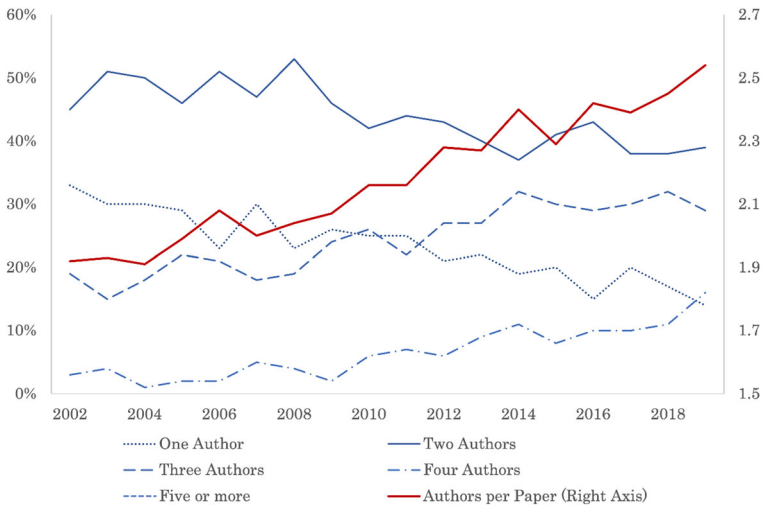
Fig. 2 Number of authors of published papers in T5

name. We rely on three different databases: the first-names' database published by the USA. Social Security Administration, created using data from Social Security card applications; the database constructed by Tang et al. (2011), who use Facebook to collect data on first names and self-reported gender; and finally, the names' database developed by Bagues and Campa (2017). We check manually any candidate who (a) falls within the [0.05 0.95] probability interval of being male/female or (b) cannot be found in any of the databases.

We convert the original sample of articles into an articles-authors sample. We transform the original 5311 articles to a total sample of 11,721 (with implied 9840 articles-men authors, and 1881 articles-women authors). Except otherwise indicated all measures below are computed over this augmented articles-authors sample.

Figure 3 depicts the share of female authors (right axis), which has been increasing (with fluctuations) at a rate of 6.2% per year, (compared to men's share average rate at 3.7%), reaching 20% share during a couple of years in the recent past. Despite female authors are increasing at a higher rate, and that there have been an important improvement in the last decades, women are clearly under-represented in T5 publications. These data are consistent with the data from the report of the Committee on the Status of Women in the Economics Profession, Chevalier (2020). Figure 4 compares the evolution of the share of women in the different professor categories of the top 20 Schools of Economics in the USA in 2020 with the proportion of female authors in top 5. Notice that the share of female authors is very similar to the 20,4% share of women in the faculty of the top 20 Schools in the USA on 2020. In line with Heckman and Moktan (2020), the rate of increase of female coauthors in T5 seems to parallel the rate of increase of female full Professors in these departments. The average proportion of females that are full professor in Spain and the EU average are very similar as well.

We have split the description of the data into two figures: one for single gender groups and another for mixed teams. Figure 5a shows the corresponding co-authorships
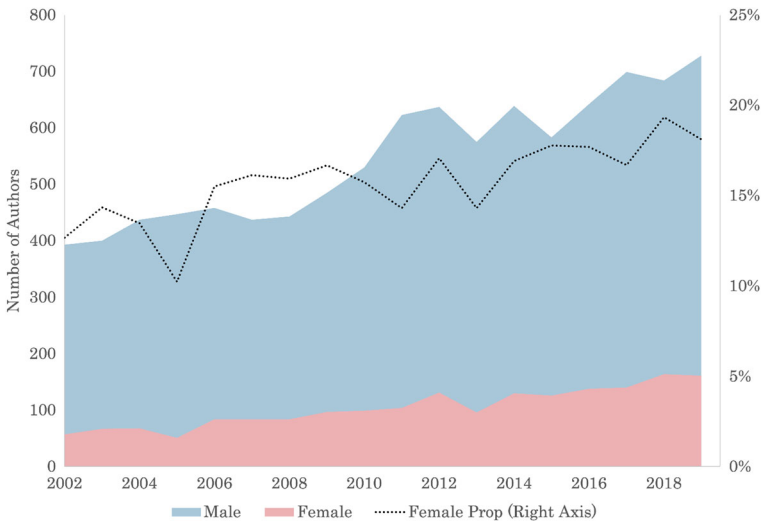
Fig. 3 Number of article-author observations by gender and the share of female articles
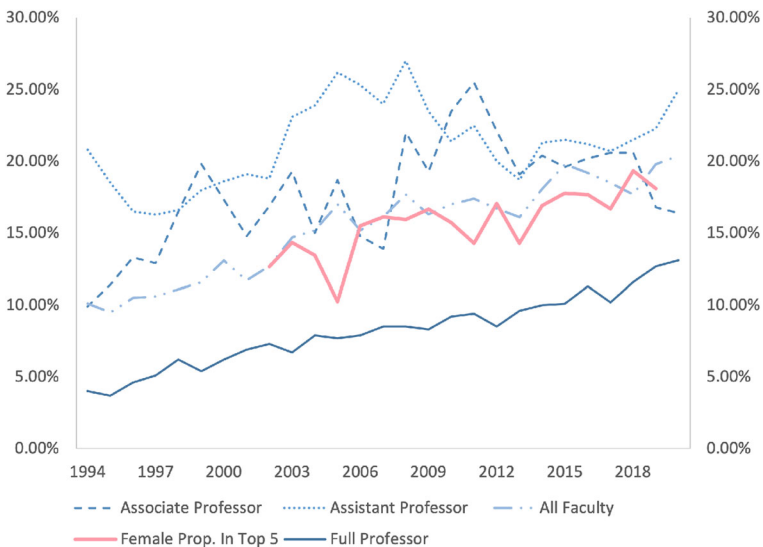

Fig. 4 The pipeline for top 20 economics departments: percent and numbers of faculty and students who are women. *Source* CSWEP Report, 2020 and own elaboration

pattern when the set of co-authors are single gender groups. The more salient feature of these data is that while the share of sole male authors has been declining from 30% of total, to slightly above 10%, the share of sole female authors has been stable over the entire sample, at a share close to 5%. We want also to point out that despite the slow decline, two males are the most common co-authors team.
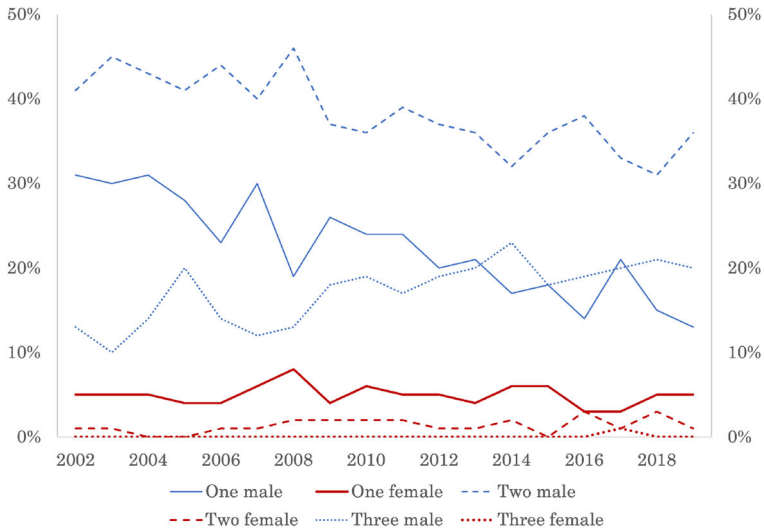
(a) Percentage of T5 articles coauthored by single gender teams.
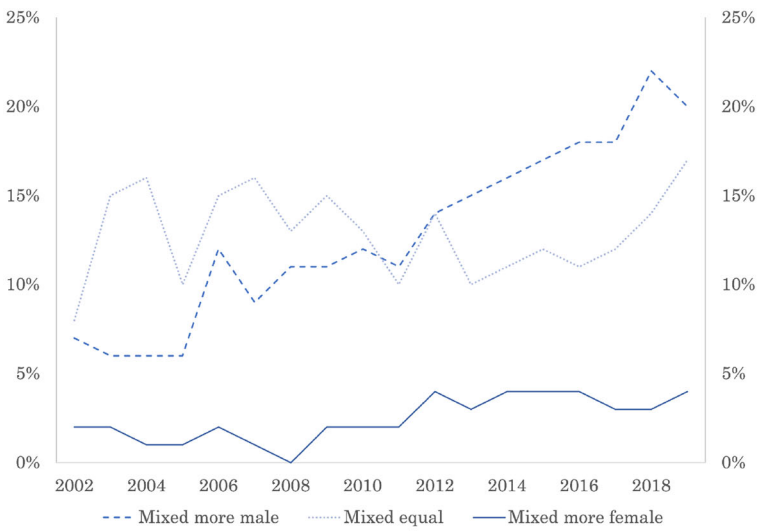
(b) Percentage of T5 articles coauthored mixed gender teams.

Fig. 5 Co-authorships patterns in T5 journals

The equal share of male-female authors has been fairly stable at about 12% (92.7% of these articles are, in particular, one male-one female). Alternatively, the share of articles with at least one woman and at least two men has been increasing from nearly 5% over total to around 14%. Thus, the strongest trend in data seems to be associated with the participation of female authors in articles with relatively more male authors.
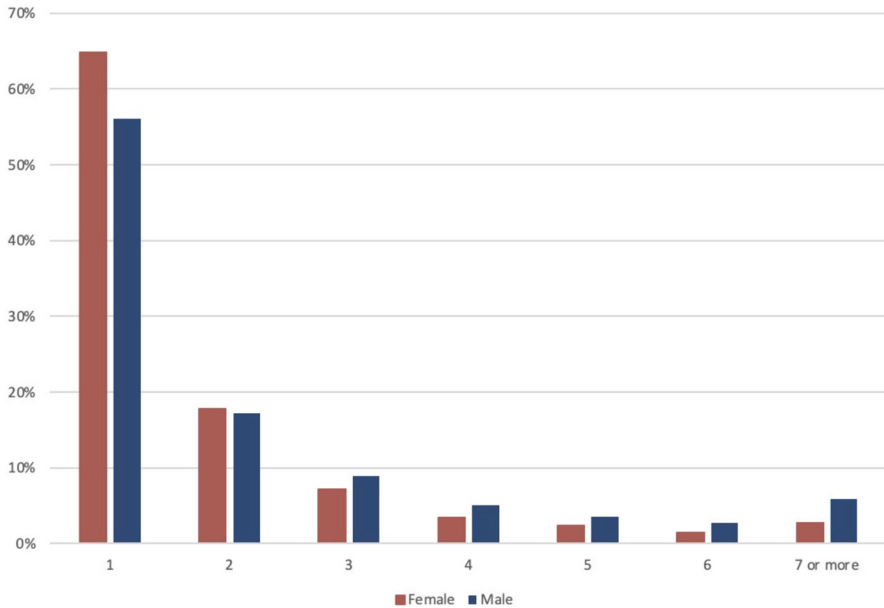
Fig. 6 Distribution of number of T5 papers published by gender

Figure 6 shows the distribution of the number of published papers by gender. Conditioning on having published in T5 journals, females are more likely than males to publish only one or two papers, while the proportion of authors that have published more than three papers is greater for males than for females. Clearly though, more than 80% of either female (15% of the distribution) or male authors have published less than two T5 over the last 20 years. This is an important fact for understanding the role of superstars in the profession as well as the mechanisms underlying the formation of networks of coauthors.

3 The Empirical Model: Structural Topic Model (STM)

Our empirical strategy is to use unsupervised machine learning techniques to uncover the hidden structure of our text documents.[6] By unsupervised we denote the absence of human intervention in order to identify the latent topics behind the abstracts of articles published in the T5 journals during the period 2002–2019. For us, an abstract is a set of words and these words have different probabilities to belong to one or several latent topics. Informally, when we are writing on a particular topic there are words that are used more often than others. Our objective is to provide a low-dimensional representation (topics) of a high-dimensional object (abstracts) while retaining as much as possible its informational content.

---

[6] For an excellent non-technical introduction to machine learning, see Hansen et al. (2018).

The baseline for topic modeling is the LDA algorithm (latent Dirichlet allocation) developed by Blei et al. (2003) and also the most popular machine learning algorithm in reducing the dimensionality of text documents.[7] In this paper, we use an algorithm called STM (structural topic model) developed by Roberts et al. (2019), which can be understood as a refinement for this LDA algorithm. This topic model is said to be structural because it allows the use of "covariates" to inform about the structure (partial pooling of parameters). These covariates in our case are going to be the different journal names and the different years in the sample. The idea is to better capture along these dimensions the changing relationship between words in abstracts and the latent topics. Next, we want to explain the algorithm and the outcome variables, and in "Appendix A" we provide a more technical discussion over STM and LDA.

We start by describing the inputs. We have our 5311 abstracts (or documents) to extract all the words. First, we have to "clean" this set of words in order to reduce the vocabulary and select terms with more informational content. This helps us for a better estimation of more semantically meaningful topics. The corpora is the set of unique words that we obtain, after converting to lower case and remove from the original raw text common stopwords,[8] as "for" or "in." Also, we prune the words until we get their original linguistic root ("educ" instead of "education") and eliminate the words that appears one or two times only.[9] In our case, we start with a set of 13,835 different terms and end up in a corpora of 4241 of unique words.

The second step is to represent our text data in a document-term matrix of D rows (5311 abstracts) and V columns (4182 unique words in our corpus) where the element (d, v) of the matrix is the number of times the v_{th} unique word appears in the d_{th} abstract. This document-term matrix that reduces the dimensionality of our original text variables is the input of the algorithm. Our objective is to find a probabilistic topic model that is able to explain the document-term-matrix in two additional steps. First by identifying K topics in our corpora and then by representing documents as a combination of those topics. What is a topic? The topic k is a probability distribution β_k over all the unique words of our corpus, where β_k^v is the probability that topic k generates word v. Each document d has its own distribution over the set of topics θ_d. This captures that each document/abstract can refer to several topics. Then, θ_d^k would mean the weight of topic k in document d. The probabilistic topic model is described by these topic β_k and document θ_d distributions. Given that, we can compute the probability that an arbitrary word in the document d coincides with the v_{th} term is $p_{dv} = \sum_k \beta_k^v \theta_d^k$. Using these probabilities, we can obtain the total likelihood of our data, $\prod_d \prod_v p_{d,v}^{n_{d,v}}$, where the $n_{d,v}$ corresponds to the elements in the document-term matrix (the number of times the v_{th} unique word appears in the d_{th} abstract).[10]

This total likelihood is our "objective" function. In a nutshell, the LDA and the STM algorithms are designed for finding numerically the stochastic model of latent topics (the distributions β_k and θ_d) that better suit our document-term matrix, that is

---

[7] For technical description of the LDA algorithm, see the original article of Blei et al. (2003) and also Hansen et al. (2018) that is the first paper that uses the LDA algorithm in the economic literature.

[8] In particular, we remove the stopwords from the SMART list, developed at Cornell University in 1960.

[9] See "Appendix B" for the details of this pre-processing.

[10] See Hansen et al. (2018) for a precise description of the computation of the total likelihood.

that maximizes this total likelihood. We are going to skip here further details on the algorithms we use, and we refer the interested reader to "Appendix A" (and also to Roberts et al. 2014). However, we want to make two important observations.

First, as indicated above, we are implementing STM instead of LDA. The main advantage of STM for our data is that we can use very relevant covariate information about our documents in order to improve parameter estimation.[11] In particular, for each document/abstract we interact the year of publication as well as the journal name. We take advantage of the variability of the abstract along the time and across journals for improving the estimation of our stochastic model in particular of the distribution θ_d).

The second important observation refers to the determination of the number of topics. We can follow two strategies. One, it is to find the number of topics that better fits the data, which usually leads to a large (optimal) K. The alternative is to force the algorithm to use a given number of topics for facilitating the interpretation of those. For our baseline analysis, we use the first approach and we work with 54 topics, but we also pursue the estimation of our stochastic model using a fixed number of topics to facilitate comparison with the results in existing literature.

Previous literature, using JEL codes (for example, in Card et al. 2019) or research areas in top departments (for example, in Dolado et al. 2012) have concentrated in a broad definition of topics as fields of research, say, Labor or Econometrics. However, the unsupervised learning methodology we use allows us to go beyond pre-labeled research areas so as to capture more subtle differences, such as writing style, particular methodologies, or the variation in research questions. For example, our methodology allow us, when identifying latent topics, to separate two papers of labor economics, but one more applied and other with a theoretical contribution. We consider our approach a promising tool to analyze if there are horizontal gender differences in economics research, that is, whether or not male and female write different articles even within the same research field. For this reason, in the next section we will analyze our stochastic model with $K = 54$ topics, while in Sect. 5, we will be focusing on estimating our stochastic model with $K = 15$ topics. In addition to these two exercises, in Appendix we extend our original sample for including the abstracts of 1117 articles published as Papers and Proceeding in *AER*, between 2011 and 2018 (before 2011 these types of papers do not have abstracts and after 2018 are published in a different journal). We will show that for this extended sample the optimal number of topics increases to $K = 70$. While we have preferred to exclude these papers of the main baseline analysis because these are very short papers with very different editorial processes than regular submissions, this extended sample generates interesting new insights.

4 Gender Differences in Latent Estimated Topics

As we said above, the number of topics that best fits the text data is 54.[12] We estimate probabilities for each document to belong to this set of built-in latent topics using the

---

[11] In Cabrales et al. (2018) there is an attempt to impute also gender as an additional covariate for the articles published in the British press by looking for female names in the body text of this articles.

[12] In "Appendix C", we provide a formal discussion about the optimal number of topics.

structural topic model. The STM output is summarized by the latent topics displayed in Fig. 7 that shows the key words associated with each of the 54 topics. The words within each row are ordered left to right by the probability they appear in each latent topic. Eventually, we could assign some labels to latent topics, based on well-known fields names in economics. For instance, we can associate the more prevalent topic in the sample in expectation, topic 28, to international trade. Likewise, the second more prevalent topic in the distribution, topic 9, may be associated with Econometric Theory. However, this is not the goal of the analysis as we have indicated above. The important thing is that latent topics may be related to something beyond research fields, as methodology or style of writing. These latent characteristics hide gender differences too.

4.1 Topic Prevalence

Once we have identified the estimated latent topics, we can analyze how our documents/abstracts are distributed among them. In allocating an abstract to a particular topic, we consider our underlying θ_d distribution. Then, we assign document d to different topics with different probability weights. Following this approach, Fig. 8 shows latent estimated topics in a way that also illustrates the number of documents in each topic, notice that in Fig. 8 the size of the circle is proportional to the expected number of documents in the topic (we have also reproduced numerically this information in a column in Fig. 7). As we cannot make a mapping of our 54 topics to particular fields of research, it is difficult to interpret the information of Fig. 8 regarding the size of the topics. For example, topics 11, 9 and 21, in Fig. 8 are related to "Econometric Theory," and are relatively large compared with other topics. However, if the algorithm would have introduced more topics within "Econometric Theory," each topic would have had a smaller mass, the weight of the research field being the same. In other words, our perception of the successful topics is affected by how the research field is split into topics.

Figure 8 also contains information over the connectedness between topics. For example, if the latent topic k is closer to k' than k'', it means that the distribution β_k is more alike to the distribution $\beta_{k'}$ than to distribution $\beta_{k''}$. Looking at Fig. 7 and the description of the latent topics in Fig. 8, some interesting patterns arise. For example, the previous discussed topics 11, 9 and 21 ("Econometric Theory") are in someway isolated from the rest of topics. In Fig. 8, we can also identify some other clusters of topics, for example (east in Fig. 8) 51, 34, 23, 2, etc., are topics related to Macro-Finance, closer to those in Econometric Theory, but not that much; (west in Fig. 8) 50 is a central node of a set of topics related to Political Economy and Institutions); (southwest in Fig. 8) 29, 32, 22, etc., are topics related to microeconomics (contract theory, decision theory, etc.). Finally, applied areas as labor, international-development, or public economics are located around topics 19, 49, 28, and 48 (north in Fig. 8). In "Appendix D", we undertake a more formal analysis of the distance between topics using a simple correspondence analysis of the probability matrix for documents to belong to the different latent topics. We find the corpus organized along

Topic										Topic Prop.	Female Prop.
Topic 28	trade	countri	product	export	intern	import	firm	sector	factor	develop	3.8%	17.8%
Topic 9	estim	method	sampl	data	asymptot	paramet	consist	use	error	bias	3.5%	13.4%
Topic 11	condit	variabl	function	identif	identifi	restrict	estim	distribut	instrument	bound	3.3%	15.5%
Topic 29	experi	subject	experimen	behavior	treatment	predict	learn	evid	theori	differ	2.8%	17.5%
Topic 22	prefer	choic	decis	util	individu	make	altern	behavior	set	maker	2.7%	14.1%
Topic 21	test	statist	asymptot	distribut	method	paramet	confid	propos	forecast	bootstrap	2.7%	15%
Topic 19	school	student	effect	educ	colleg	score	test	teacher	program	assign	2.6%	17.8%
Topic 48	wage	worker	employ	firm	product	job	increas	labor	plant	skill	2.6%	15.4%
Topic 37	equilibrium	dynam	general	equilibria	exist	economi	condit	stochast	solut	uniqu	2.5%	10.8%
Topic 51	shock	polici	monetari	inflat	aggreg	respons	money	real	nomin	volatil	2.4%	13.2%
Topic 16	belief	agent	expect	prior	ration	probabl	signal	util	set	learn	2.3%	10.1%
Topic 6	game	player	strategi	payoff	equilibrium	play	bargain	repeat	cooper	equilibria	2.3%	10.4%
Topic 2	price	cost	adjust	chang	data	firm	demand	good	markup	relat	2.3%	17%
Topic 49	women	children	parent	femal	men	famili	educ	marriag	child	birth	2.2%	32.8%
Topic 53	market	inform	trade	price	asset	valu	trader	privat	advers	select	2.2%	15.1%
Topic 15	welfar	cost	benefit	insur	gain	polici	estim	loss	reduc	use	2.2%	18.7%
Topic 33	return	firm	stock	manag	asset	equiti	investor	portfolio	predict	size	2.2%	18.4%
Topic 32	contract	agent	princip	commit	optim	hazard	incent	moral	inform	problem	2.1%	11.6%
Topic 50	polici	polit	govern	parti	elect	voter	power	politician	elector	public	2%	13.5%
Topic 34	financi	invest	constraint	recess	shock	asset	firm	aggreg	credit	financ	2%	15.6%
Topic 3	risk	avers	consumpt	ambigu	util	discount	prefer	expect	asset	intertempor	2%	14.5%
Topic 47	consum	firm	product	demand	market	good	price	profit	advertis	competit	1.9%	15.1%
Topic 41	percent	health	insur	increas	hospit	estim	care	patient	drug	use	1.9%	22%
Topic 18	region	econom	area	local	growth	land	agricultur	locat	develop	data	1.9%	14.4%
Topic 43	household	hous	consumpt	spend	incom	expenditur	increas	effect	respons	data	1.8%	15.1%
Topic 45	cycl	busi	product	industri	fluctuat	chang	demand	volatil	aggreg	entri	1.8%	14.4%
Topic 40	optim	alloc	effici	distort	economi	privat	condit	ineffici	resourc	polici	1.8%	14.1%
Topic 27	incom	earn	inequ	data	differ	measur	survey	distribut	use	mobil	1.7%	17.4%
Topic 52	capit	human	invest	skill	growth	accumul	differ	labor	account	life	1.7%	14.5%
Topic 26	market	match	stabl	friction	competit	labor	agent	labour	side	type	1.7%	15.6%
Topic 25	technolog	innov	product	new	firm	patent	research	adopt	knowledg	spillov	1.7%	19.5%
Topic 44	inform	coordin	action	communic	strategi	game	outcom	sender	signal		1.6%	14.9%
Topic 10	mechan	implement	incent	transfer	type	design	compat	post	agent	problem	1.6%	10.9%
Topic 5	auction	bid	bidder	buyer	seller	valu	price	revenu	privat	inform	1.6%	14.8%
Topic 4	state	unit	right	issu	econom	protect	problem	institut	properti	resourc	1.5%	15.3%
Topic 12	social	network	individu	incent	interact	opportun	depend	connect	link	secur	1.5%	19.9%
Topic 17	bank	credit	polici	fund	crisi	lend	liquid	loan	financi	market	1.5%	14.5%
Topic 42	public	regul	enforc	good	privat	law	provis	punish	legal	cost	1.5%	18%
Topic 13	work	program	labor	suppli	hour	increas	transfer	time	particip	home	1.4%	18%
Topic 20	tax	reform	incom	rate	increas	taxat	margin	chang	optim	effect	1.4%	16.9%
Topic 23	debt	default	borrow	govern	credit	bond	fiscal	sovereign	market	matur	1.4%	16.8%
Topic 1	econom	studi	name	correct	bias	black	measur	data	signific	racial	1.3%	18.7%
Topic 30	firm	contract	ownership	vertic	integr	adopt	industri	cost	supplier	exclus	1.3%	21.8%
Topic 38	group	ethnic	member	trust	evid	segreg	countri	increas	cultur	chang	1.3%	19.8%
Topic 36	inform	vote	signal	voter	aggreg	bias	privat	strategi	elect	larg	1.3%	15%
Topic 39	rate	exchang	interest	currenc	countri	real	patient	donor	regim	transplant	1.2%	13.6%
Topic 31	save	citi	retir	account	popul	life	increas	german	individu	rate	1.2%	19.4%
Topic 7	vote	news	voter	media	candid	elect	estim	committe	newspap	bias	1.2%	17.5%
Topic 8	search	unemploy	worker	job	distribut	durat	wage	rate	employ	benefit	1.2%	14.7%
Topic 35	conflict	increas	violenc	crime	war	polic	outsid	option	effect	attack	1.1%	17.1%
Topic 14	rule	demand	set	ration	problem	yield	constitut	optim	function	util	1%	10.5%
Topic 46	project	effort	team	perform	redistribut	outcom	win	competit	one	prize	0.9%	19.4%
Topic 24	qualiti	delay	probabl	accept	fee	order	card	offer	paper	higher	0.8%	14.8%
Topic 54	import	use	addit	data	sever	relat	support	analys	find	limit	0.3%	15.7%

Word Prevalence (%)

5   10   15   20

Topic Proportions (%)
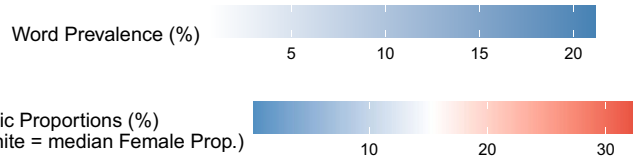(White = median Female Prop.)

10   20   30

Fig. 7 Optimal *K* topics ranked by prevalence in the corpus

two dimensions: Dimension 1 can be interpreted as going from Applies to Theory, whereas Dimension 2 goes from, say, Economics to Econometrics.

Using our classification of authors' names by gender and the allocation of documents to latent topics, we can build up a similar figure with information about the gender distribution. Figure 9 shows latent topics where the sizes of circles are proportional to the percentage of female authors working in such topics (we have also reproduced numerically this information in the last column in Fig. 7).
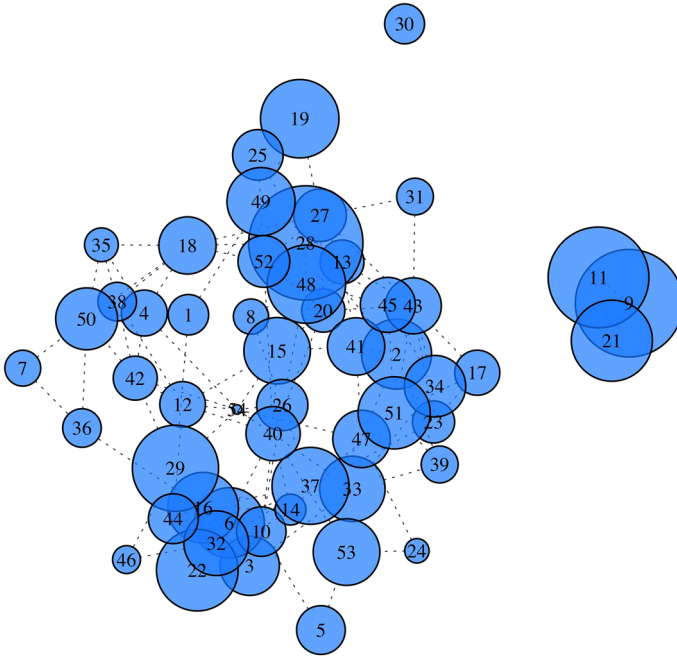
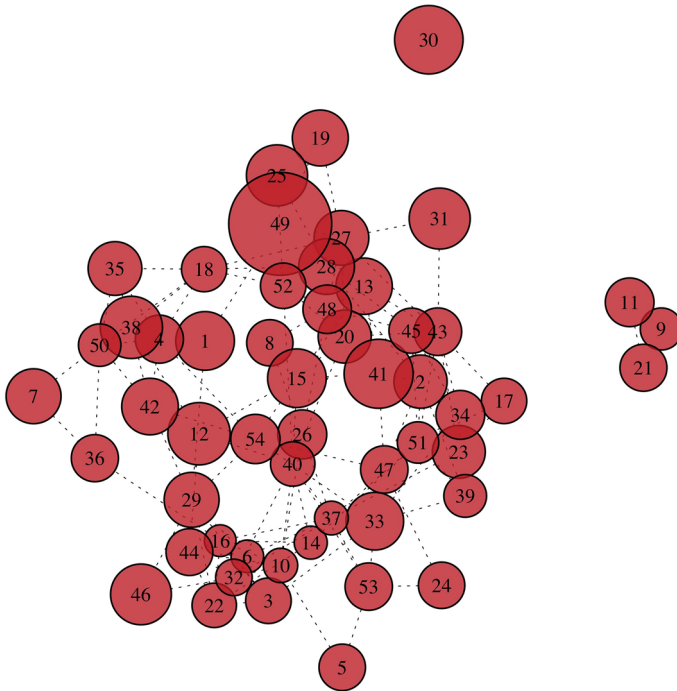**Fig. 8** Connectedness between topics and the fraction documents/abstracts in each topic (θ_d distribution)

Fig. 9 Connectedness between topics and the female authors documents/abstracts in each topic

(a) Topic 49 (highest prop. of female authors).   (b) Topic 16 (lowest prop. of female authors).

Fig. 10 Topic Word Clouds: Topic 49 vs Topic 16

Fig. 11 On the presence of women, by topic: mean and one standard deviation across time

Figure 9 provides interesting evidence of the main message of this paper, male and female display different patterns when doing research. Independently of the grade of under-representation of women in the profession, if there were not significant gender horizontal differences we would expect that sizes of latent topics measure for the proportion of females were similar. On the contrary, we observe an uneven distribution of sizes.

There is a small subset of topics (north in Fig. 9), specially topic 49, with a relative high proportion of females, that moreover seem to be closely connected (according to the terminology for applied economics fields). On the contrary, there is other set of topics (for example, southwest in Fig. 9) that are also closely connected and where the presence of females is scarce (around terms common to economic theory research questions).

4.2 Topic Analysis and the Gender Distribution

As we said above, it is difficult to describe the precise semantic meaning of the latent topics when we are working with $K = 54$. We are able, however, to look closer to the latent topics where females are more or less prevalent and its potential implications. In particular, Fig. 10 shows that the latent topic with the highest proportion of female authors is topic 49 (32.8% as indicated in Fig. 7). On the contrary topic 16 turns out to be the topic with the lowest proportion of females (10.1% as indicated in Fig. 7). As a simple illustration, Fig. 10 represents these topics as word clouds, where the size of terms in the cloud is equivalent to its probability in the latent topic distribution β_k.

The words that seem to be more prominent in the cloud 49 are women, men, parent, children, health, etc. These words could be easily linked to research fields, as gender or health economics, traditionally associated with women. Similarly, the word cloud of topic 16 seems to be related to Micro theory that has been often labeled (while not statistically) as an area where there are less female than average.

Latent topics may differ in other dimensions beside semantic content. For instance, Hengel (2020) uses readability scores to measure the quality of writing of article abstracts.[13] We have implemented E. Hengel's `Python` module `Textatistic` to compute readability results over the article abstracts across our latent topics. The finding is that scores across more female topics are better rated than across more male topics. However, it is hard to disentangle the role of the prevalence of female authors face to face the wording within a topic. Moreover, scores that are outliers should be properly treated to ease comparisons. We leave the study of these readability issues implying fundamental gender differences for further research.

Rather, Fig. 11 shows the mean of the presence of women authors by topic, together with the standard deviation of this presence over the sample of years. For some latent topics, the proportion of females is larger than the average (which is 15.9% over the period 2002–2019), reaching a proportion of 33% for topic 49. On the contrary, females are specially underrepresented in other topics, as topic 16, with only a 10%. Dispersion over time differs also across topics, and it seems that is higher for topics with higher proportion of females (the correlation between dispersion and the proportion of females is 0.35). While it is true that the proportion of female authors has been increasing in the last two decades from around 13% on 2002 to 19% on 2019, we do not see a trend in the dispersion of the proportion of females by topic. Consequently we see the prevalence of females across topics as a signal of gender "horizontal" differences in research.

Nevertheless, for having a more accurate picture of this "horizontal" differences, we need to add the information regarding the relative prevalence of the topics. It could be possible that females are unrepresented in a particular topic, and this circumstance having little impact as far as this topic contains very few published papers.

Figure 12 shows the distribution between males and females across topics normalized for having the same size. This gives us the propensity that, say, a female authored paper belongs to any of the 54 topics. We rank the topics according to probability of

---

[13] As E. Hengel discusses in detail, abstract readability is strongly positively correlated with the readability of other sections of a paper.
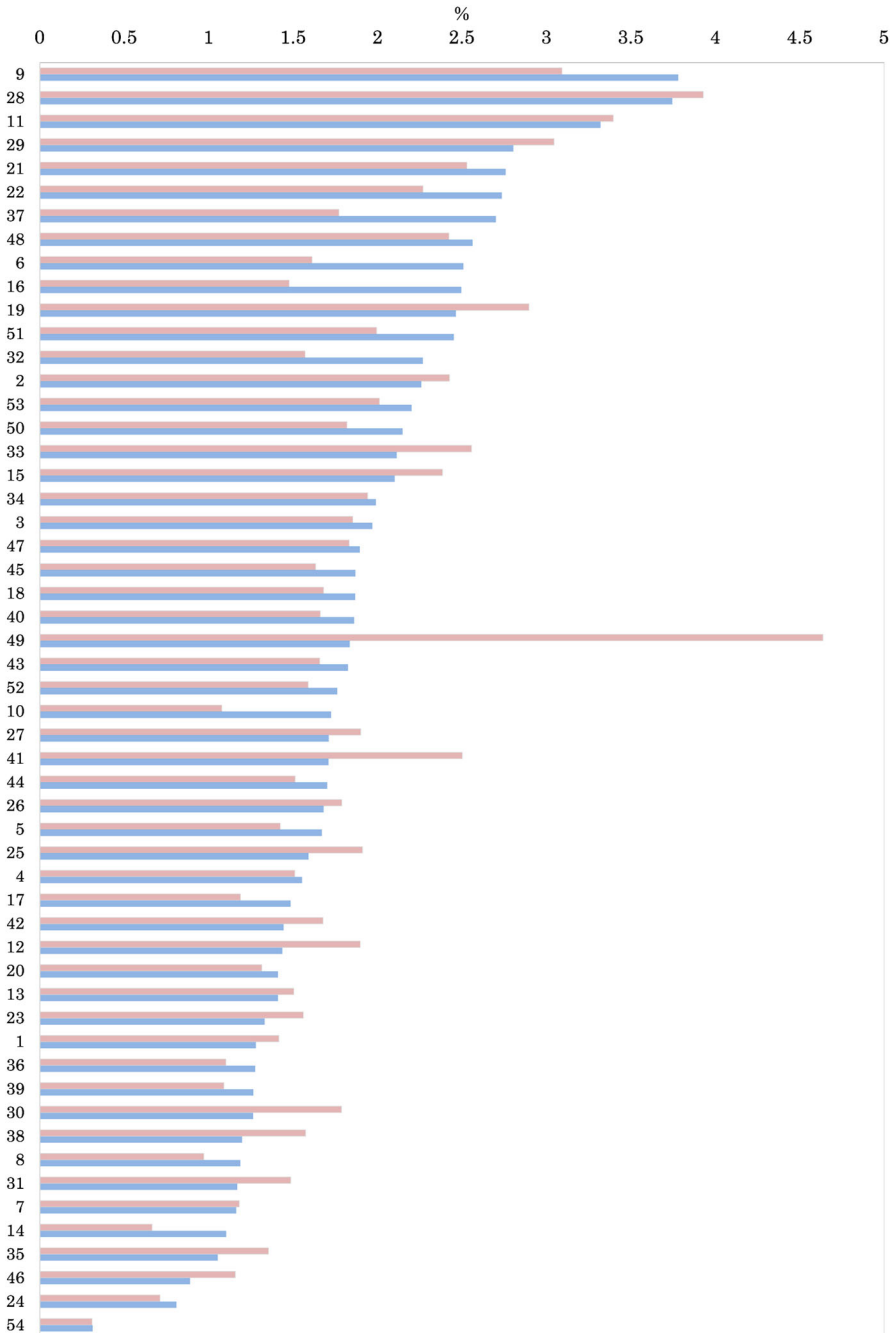
Fig. 12 Empirical distributions across topics between males and females (conditional of having published an article in Top 5)
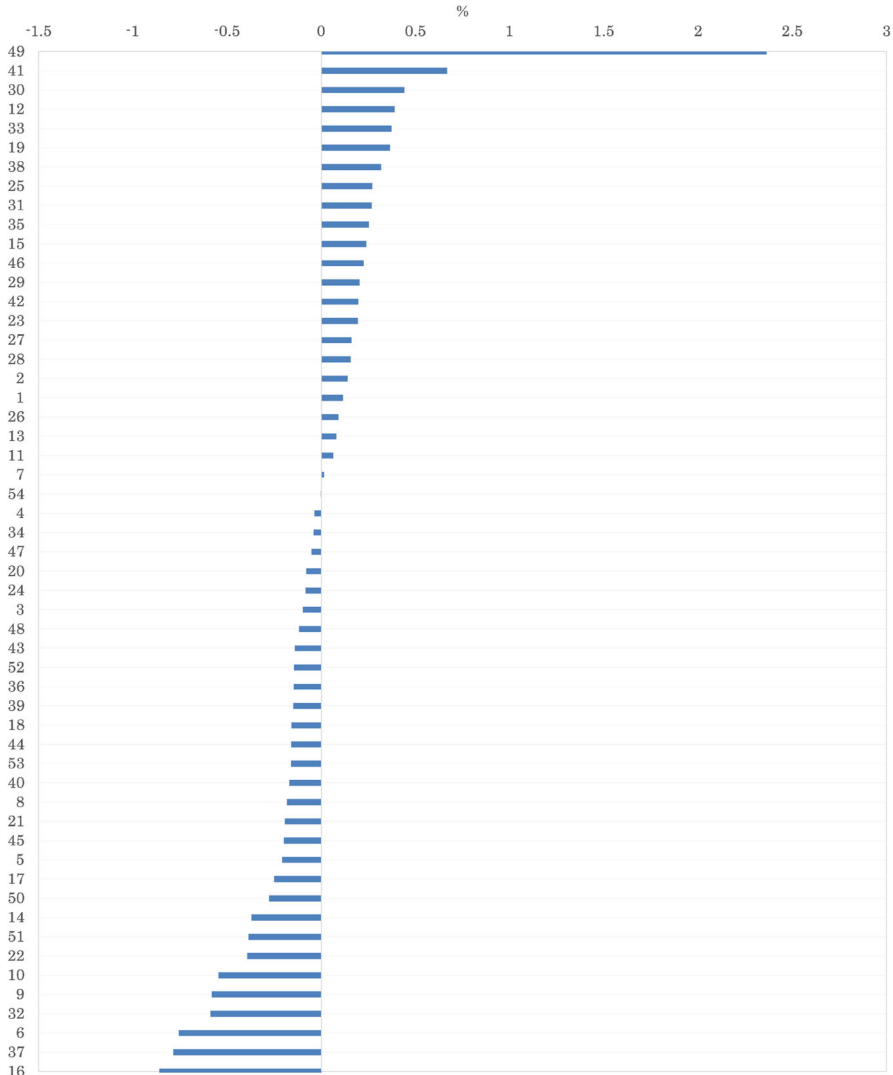
Fig. 13 Relative propensity of publishing papers by females over topics

being chosen by a male author. This figure provides evidence that male and female authors either have different preferences or follow different strategies when pursuing and publishing their research. We observe that topics with higher "demand" by males are also highly demanded by females. However, there is a set of topics, for which the proportion of published papers for men are high, which are less attractive (o more difficult to publish) for females. In general, male and female distributions are different, with the salient feature of topic 49 for females, that it is a clear spike in the female distribution of published papers.
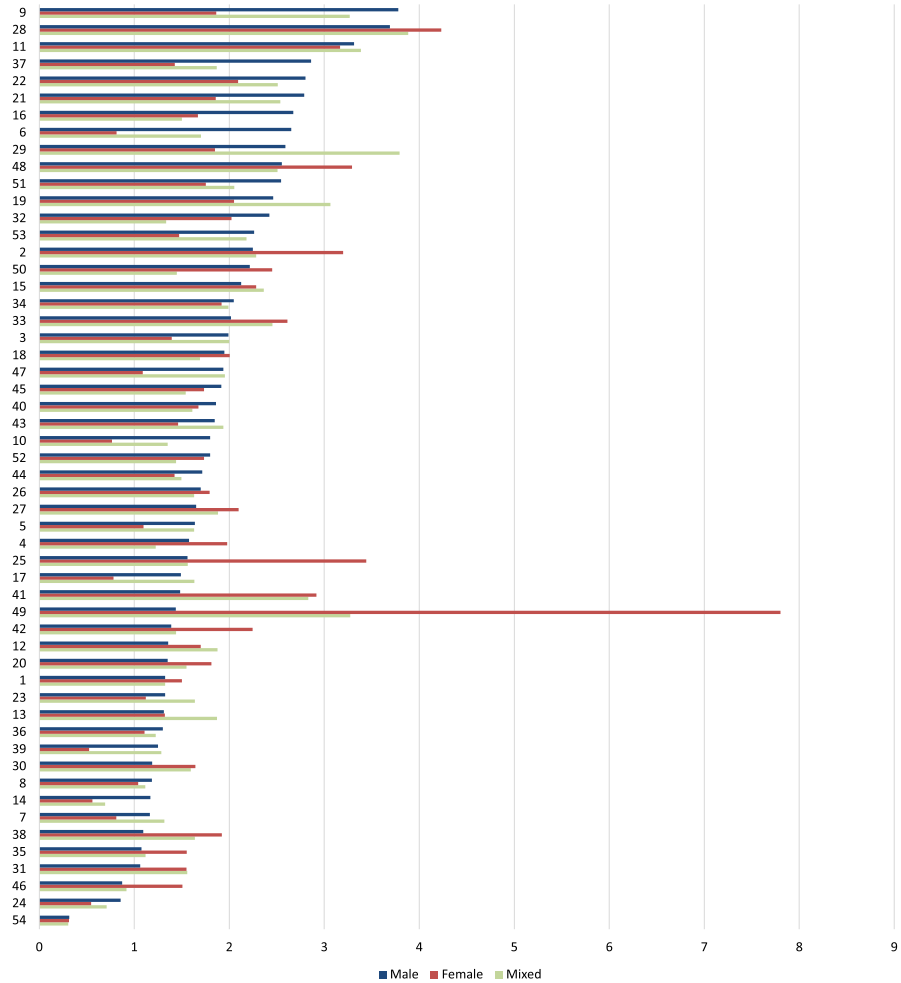
Fig. 14 Empirical distributions across topics between males, females and mixed authorship (conditional of having published an article in top 5)

We confirm this evidence with a complementary Fig. 13 representing the dispersion of published female authored papers across topics, but accounting also for the prevalence of latent topics. In particular, for each topic we have the proportion of published papers by female authors (taken from Fig. 12) minus the proportion of published papers in this topic overall. Conditioning on having published a paper, male and female would be equally likely to publish a paper in a specific topic, this difference would be zero. Then, we can interpret this difference as the excess propensity to publish a paper in a particular topic by females. These differences can be positive or negative, and the sum over all topics is zero. The figure shows that there are topics for which the propensity of publishing papers by females is higher than males, and the opposite.
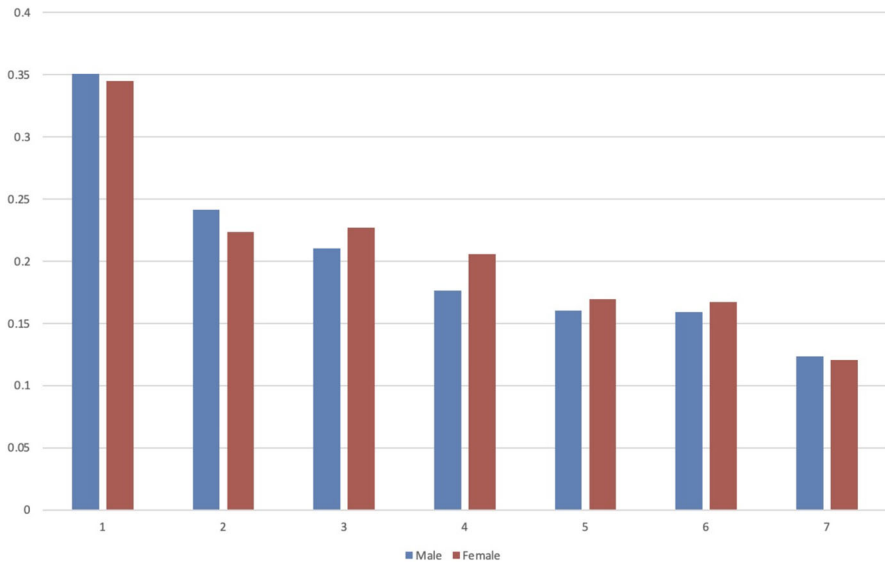
Fig. 15 Diversify across latent topics by gender (HHI)

Again topic 49 but also topics 41 (health) and 30 (applied IO) are in one side, while theory topics as 16 or 37 are in the other side.

In order to analyze the pattern of coauthor-ships we have pooled the articles in three groups, papers written by male authors, by female authors, and gender mixed team of authors. The main results are summarized in Fig. 14 that shows that there is a important difference between the pattern of latent topics between sole male teams and sole female teams, while mixed teams generate an intermediate distribution over the latent topics.

Finally, we want to address a related but different question, how male and female diversify across topics. For example, when writing an article, an author may contribute to a single latent topic or several, authors that have published several papers may have written similar articles or they could have been more diverse: are these diversification patterns different for males and females? For addressing this question, the first step is to choose a measure of latent topic dispersion/concentration. A natural candidate is the Herfindahl–Hirschman Index (HHI) that is used to measure the concentration in a market.

The HHI index is calculated by squaring the market share of the firm (the topic) that compete in a single market and then summing up the resulting numbers $\text{HHI} = \sum_{i=1}^{N} s_i^2$. We apply this index to our problem as follows. For each author (the market), we identify all the latent topics that she has contributed to (the firms). For each article the algorithm computes a probability distribution over the latent topics. We repeat the process for all articles of the same author. Then, the cumulative probability divided by the number of articles is the contribution of the author to this particular latent topic (the market share, s_i). For example, if an author publishes very similar papers related to a single or a few latent topics, her HHI will be high. On the contrary, authors with a more

diverse research agenda will have a lower HHI. Figure 15 shows the corresponding average HHI for males and females.

We have computed the HHI controlling for the number of papers by author. It is clear that an author that has published more papers is likely to have contributed to a larger set of latent topics and therefore she must have a lower HHI. Interestingly, the figure shows some differences between genders in terms of diversification. Females are more diverse (lower HHI) when publishing one or two papers, but less (higher HHI) when publishing a larger number of papers in the Top 5.[14]

5 Topics as Research Fields

In this section, we estimate the stochastic model with a lower number of topics, with two objectives. On the one hand, a low K facilitates the semantic interpretation of topics and then to analyze, for instance, whether or not, the weight of a particular field in the T5 has increased over time. On the other hand, a low number of topics will allow us to frame our results with previous literature that has used a small number of categories linked to JEL codes and research areas in top departments. After estimating the model for a range of $K \in 10, \ldots, 20$, we have found that $K = 15$ is a number of topics for which the estimated model performs better in terms of fitting to the data and the semantic content of the latent topics at the same time. The model with $K = 15$ latent topics is summarized in Fig. 16.

The reader may then wonder what additional information is contained in the unrestricted version of the structural topic model (STM). One way to illustrate on the importance of an adequate selection of the number of topics is to explore in detail the composition effects we already discussed above. We proceed as follows. First, we consider the stem "labor," and we look for it among the fifteen more frequent words within the restricted version of the STM, that is, the version with just 15 latent topics ($K = 15$). We only find that particular word under the required frequency within topic 8 in Fig. 16. Figure 17 depicts the word cloud for that topic 8 in the restricted version of the model with $K = 15$. Clearly, in this particular case, one may say this cloud describes well the research field corresponding to JEL code *J*, which is Labor and Demographic Economics.

The key idea with the structural topic model is that a field like "Labor" can fit many research lines in the unrestricted version of the model, in our case the one with 54 latent topics. When we look for the stem "labor" within the 54 latent topics, we find it among the fifteen more frequent words in as many as six topics. Figure 18 illustrates on the most prevalent among these topics which are: Labor Search, Labor Supply, Human Capital, or Productivity Analysis. Notice, in particular, that there are important differences on the prevalence of females across these different subtopics, from 18 per cent in the more policy oriented topic which is "labor supply" to 14 per cent in the more theoretical "labor search" (go back to Fig. 7 for these shares).

---

[14] The HHI is a first approximation as measure of research diversification. In the future, we want to improve the measure by taking into in account that some latent topics are close to others.

										Topic Prop.	Female Prop.
Topic 9	estim	test	distribut	condit	paramet	method	function	use	variabl	asymptot	10.3%	14.3%
Topic 2	product	firm	trade	industri	import	cost	data	countri	sector	export	8.4%	17.3%
Topic 3	asset	financi	invest	bank	risk	return	market	credit	rate	debt	7.9%	15.4%
Topic 14	prefer	choic	decis	util	risk	expect	subject	theori	behavior	individu	7.9%	14%
Topic 6	equilibrium	game	inform	player	equilibria	payoff	action	learn	signal	belief	7.2%	11.3%
Topic 1	countri	growth	econom	incom	state	data	capit	across	citi	develop	7%	16.3%
Topic 13	increas	household	percent	insur	health	estim	hous	women	incom	children	6.9%	23.4%
Topic 5	price	market	consum	inform	competit	cost	demand	auction	good	profit	6.9%	15.3%
Topic 10	agent	contract	optim	effici	mechan	alloc	incent	problem	match	condit	6.8%	12.7%
Topic 12	social	experi	individu	group	perform	network	incent	inform	effort	manag	6.2%	18.2%
Topic 15	polici	tax	rate	welfar	monetari	optim	govern	inflat	respons	shock	5.5%	14.7%
Topic 8	wage	worker	labor	employ	market	job	earn	unemploy	increas	skill	5.3%	16.1%
Topic 7	polit	vote	voter	polici	elect	govern	parti	candid	crime	power	4.6%	16.6%
Topic 11	effect	school	treatment	program	student	test	outcom	random	assign	use	4.6%	17.6%
Topic 4	technolog	innov	invest	regul	adopt	institut	right	resourc	increas	enforc	4.5%	18.8%

Word Prevalence (%) 2 4 6 8

Topic Proportions (%)
(White = median Female Prop.) 5 10 15 20

Fig. 16 Latent topics ranked by prevalence in the corpus with $k = 15$

Important variability can be washed out when the methodology used account for the research field environment rather than for the research topic environment.

As we have anticipated, the reduction of the number of topics to $K = 15$ makes easier to label the latent topics as meaningful research fields, though. Following our previous analysis, Fig. 19a plots the latent topics showing the relative semantic distance

Fig. 17 A topic with "labor": topic 8 in the set with $K = 15$

(a) Labor Search (14% fm)

(b) Labor Supply-Policy (18% fm)

(c) Human capital (14% fm)

(d) Productivity (16% fm)

Fig. 18 Word clouds for topics with the stem "labor" among the fifteen more frequent words in the set with $K = 54$
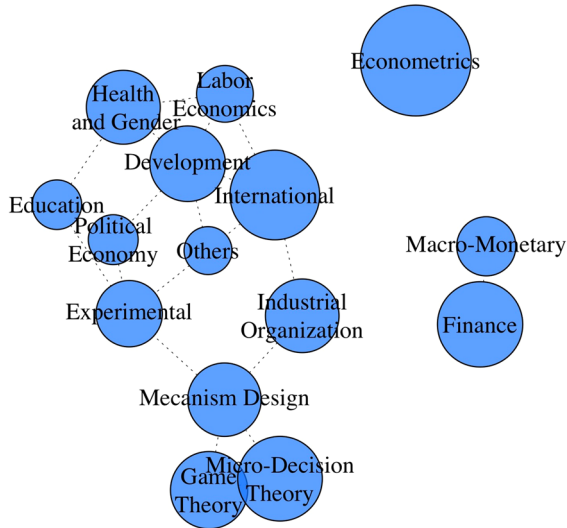
(a) Connectedness between topics and the fraction documents/abstracts in each topic (θ_d distribution).
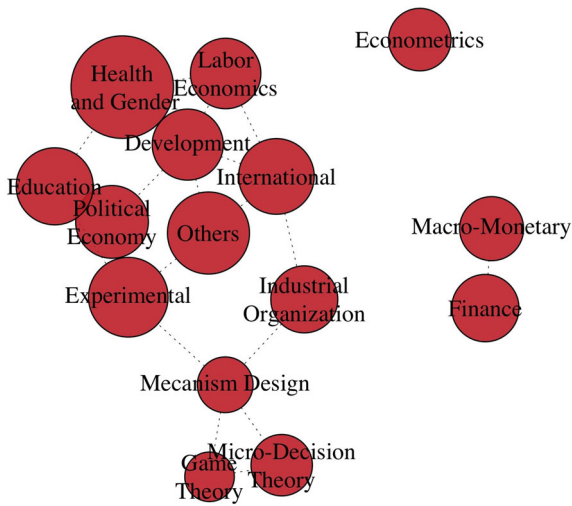
(b) Connectedness between topics and the female authors documents/abstracts in each topic.

Fig. 19 Connectedness for $K = 15$

between topics as well as their weight in terms of the fraction of documents/abstracts that they contain.

If we compare Fig. 7 (with $K = 54$) and Fig. 19a (with $K = 15$), they have a similar "geography" in terms of general areas of knowledge. Therefore, similar patterns in terms of the distances between topics arise. For example, "Econometric Theory" seems to be isolated, whereas applied fields such as Labor and Public Economics are closely connected.

Figure 19b (as Fig. 8 with $K = 54$) provides evidence of the "horizontal" differences between males and females in doing research. The results go in line with the previous literature as in Dolado et al. (2012), Chari and Goldsmith-Pinkham (2017), Beneito et al. (2021) and Lundberg and Stearns (2019) that point out that females are unevenly distributed across fields. We concur with previous literature that females are over-represented in Applied-Micro fields, specially Health-Gender, Experimental and Education and underrepresented in Econometric and Economic Theory fields, Macro-Monetary and Finance.

For example, Dolado et al. (2012) use the classification of women by research areas (JEL 20 fields) in the top 50 economic departments in 2005. The proportions they find are very similar to ours: (i) I-Health, Education and Welfare, 25%, (ii) D-Microeconomics, 14%; (iii) J-Labour and Demographic Economics, 15% or (iv) C2-Econometrics, 14.3%. In our analysis, we found that the percentage of female authors are, for example: (i) Health and Gender, 23%; (ii) Decision Theory (13.6%), Game Theory (11.4%); (iii) Macroeconomics and Monetary, 14.2%; or (iv) Econometrics, 14.4%. Having said that, the distribution of the proportion of females across these restricted topics seems to be slightly less disperse than those identified in the previous literature with other sources of data. This can be due to the fact that our methodology is more "continuous" than allocating females to fixed categories, and as far as the probabilistic model allocates females' articles to latent topics with statistical weights.

Figure 20 analyzes together the evolution of the prevalence of the topics and the proportion of females authors. For building this figure, we have computed the growth rate of topics' prevalences and topics' female proportions from the averages in the latest seven years (2013–2019) and the first seven years (2002–2008) of the sample. First, we can observe that the proportion of females have increased in all topics, but Finance (−6.6%). Regarding the prevalence, only four topics have decreased their weight in terms of prevalence, Mechanism Design (−10.3%), Econometrics (−29%), Game Theory (−22.5%) and Experimental (−8.4%). On the one hand, the topics where the percentage of women authors have risen more are Political Economy (+67.7%), Decision Theory (+42.5%), Macroeconomics and Monetary (+32.3%), Experimental (+40%) or Labor (+35%). In all of them, the women were clearly underrepresented. On the other hand, the topics where the percentage of women has grown the least, besides Finance, have been Health and Gender (+11.4%), Econometrics (+9.4%), and IO (+9.2%).

Finally, there is no clear relationship between the growth rate of topic prevalence and the increase in female prevalence. This is surprising. We do not have data about the seniority of authors, but as the proportion of female is increasing, we can expect that the proportion of females among the new entrants in the T5 market should be relatively large. New entrants should be more likely to work in "hot" topics rather
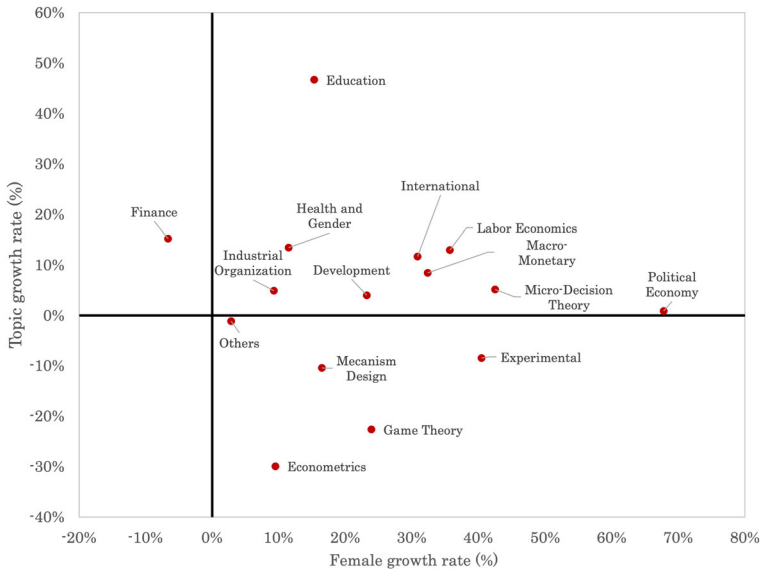
Fig. 20 Growth rates of prevalence and female proportion by topics

than in declining ones. The combination of both effects should lead to a positive correlation between the increase in the prevalence of a topic and the increase in female representation, something that we do not observe clearly in the data. However, another alternative explanation to the increase of the proportion of women in some topics is that females that already have published in top five in the past, have extended their network of male coauthors and getting more papers published.

6 Conclusions

Using unsupervised machine learning techniques and a new data base composed by the abstracts of all articles published in T5 journals in Economics for the period (2002–2019), we have shown that there are persistent and significant horizontal differences in the way males and females approach research in Economics. Using the structural topic model, we have identified latent topics for which the distribution of female authors is more uneven than with research fields. These findings are important for several reasons, because: (i) T5 publications are key for research careers and also for determining the path of economic research; (ii) the results are robust in the sense that they are automatically generated with a probabilistic model without any deterministic allocation of papers to pre-established categories or fields of research; (iii) finally, recent theoretical results by Conde-Ruiz et al. (2017, 2021) and Siniscalchi and Veronesi (2020) show that "horizontal" gender differences in the choice of research topic may lead to a gender discriminatory trap.

Beyond the scope of the present paper, we plan to extend our analysis in several directions. Firstly, we want to recollect more information about the authors, in order

to be able to capture dynamic effects. For instance, we want to differentiate between the research patterns by senior and junior authors. We want also to investigate how male and female build the network of coauthors and how this process determines the choice of latent topics. Secondly, we want to show the usefulness of the methodology and the latent topics we have identified by reviewing research questions analyzed by previous literature in academic gender gaps. For example, Hengel (2020) analyzes the differences in quality of writing of papers. She shows that female-authored manuscripts are better written and concludes that female are subject to higher writing standards. The reason might be an unwelcome gendered culture through the entire editorial process at the time of deciphering complicated texts. We are currently implementing Hengel's readability scores methodology to the latent topics. Our preliminary findings suggest that those papers belonging to topics with more prevalence of females are better written. Although this evidence can be interpreted as supporting the view that female-authored articles are better written than equivalent articles by men, it can be also the case that the results are driven by the particular topics. In other words, we need a deeper econometric analysis to disentangle if the written quality of the papers is driven by gender of the author or by the choice of the latent topics.

Likewise, Card et al. (2019) shows that female authored papers have more citations, suggesting that journals hold female-authored papers to higher standards. They have obtained this result controlling for research field. We plan to collect data on citations and review this result but controlling by latent topic. Finally, we want also to use algorithms (for example, LASSO a widely used regression analysis machine learning method) for testing if the differences between gender research patterns are important enough, for building a predictive model of gender given an observed abstract.

Declarations

Conflict of interest José Ignacio Conde-Ruiz, García, Juan-José Ganuza and Luis Puch declare that they have no conflict of interest.

Research This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix A: The Topic Model

We implement and develop the structural topic model (STM) to incorporate document-level meta-data into a probabilistic text model. The topic model is said to be *structural* because "covariates" inform about structure (partial pooling of parameters). We keep track of journal names and publication years as covariates to estimate the prevalence of topics.

The starting point to understand the STM probabilistic model is the LDA (latent Dirichlet allocation) generative model. According to LDA, the data generating process for document $d \in D$ assigns terms in vocabulary V to positions N_d in the document-term matrix, where the element (d, v) of the matrix is the number of times the v_{th} unique word appears in the d_{th} abstract. The algorithm follows the steps below

1. Draw a K-dim Dirichlet vector θ_d containing the expected fraction of words in d attributed to topic $k \in K$.
2. For each word (position) in d, sample the indicator $z_{d,n}$ from $\mathrm{Mult}_K(\theta_d, 1)$ that indicates the position n associated with a topic.
3. Sample the indicator $w_{d,n}$ from $\mathrm{Mult}_V(B_{z_{d,n}}, 1)$, where matrix B has distributions β_k over vocabulary V; $[\beta_k]$ is frequency with which terms are generated from k.

STM in its turn builds upon identifying covariates to improve the estimation of the topics. Covariates affect (i) the proportion of a d devoted to a k (topic prevalence-TP), and (ii) how much a word is used in k (topical content-TC). To this purpose:

- for TP, Dirichlet θ_d draws of document-level attention to each topic are replaced with a logistic-normal with a mean vector parameterized as a function of document covariates.
- for TC, β_k distribution is proportional to a multinomial logistic regression parameterized as indicated below.

A (partially collapsed) variational expectation–maximization algorithm is implemented to approximate the posterior (inference). Then posterior predictive checks (cf. Gelman et al. 1996) and tools for model selection as in Roberts et al. (2014) are used. Beyond TP and TC functions of document metadata, the structural topic model can be summarized as:

1. Given parameters: (i) a variance–covariance matrix for topics Σ, (ii) a matrix of observed document-level covariates X (journals names and years) and (iii) a vector γ_k (of prevalence of each topic) for each covariate,

$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2 \, I_p),$$

sample the topic proportion in each document, vector $\grave{}_d$, that is,

$$\grave{}_d \sim \mathrm{LogisticNormal}_{K-1}(\mathbf{0}' \mathbf{x}_d', \Sigma), \quad \mathbf{0} = \left[\gamma_1 | \dots | \gamma_K \right]$$

as a substitute for the Dirichlet conjugate prior, to conform the *topic prevalence model*.
2. The *core language model* given the topic proportion per document $\grave{}_d$ consists of:

- sampling the probability $\mathbf{z}_{d,n}$ that a word is in a topic: $\mathbf{z}_{d,n} \sim MN_K(\mathbf{2}_d)$, with K outcomes
- conditional on topic, choose a word from $\beta_{z_{d,n}}$, that is $\mathbf{w_{d,n}} \sim MN_V(\beta_{z_{d,n}})$, $over\,\mathbf{B} = [\beta_1|\ldots|\beta_K]$ matrix of distributions over vocabulary V.

3. The *topical content model* samples the topic word distribution $\beta_{d,k,v,}$. By now, we do not use covariates to explain topical content of documents.

Appendix B: Details of this Pre-processing Data

Pre-processing of the abstracts that conform our database is essential in order to organize the words that form the texts in an homogeneous way. The main goal of this process is to reduce the dimensionality by reducing the set of words, but at the same time trying to maximize the information contained in the words used by the authors by selecting the terms with more informational content. This helps us for a better estimation of more semantically meaningful topics.

First step is tokenization so as to differentiate words by selecting only single words (monograms), instead of bigrams, trigrams, paragraphs, etc. Then, we eliminate punctuation, and capital letters are converted to small letters. This allows as to remove duplicates, for example "Education" and "education" are different words in our database if we do not convert all the words to lowercase. Once this is done, we eliminate numbers and stopwords. By stopwords we refer to those words without any informational content: "common" words such as "and," "for" and "in." We removed the stopwords from the list SMART developed by Buckley (1985), a public list with more than 500 words. Additionally, we remove some custom stopwords that were very common in our database but not informationally relevant. These are: "download," "slides," "slide," "jel," "abstract," "paper," "author," "literature," "among," "whether," "authors," "model," "show," "showed," "shows," "find," "can," "matter," "model," "models," "may," "effect," "find," "can," "show," "paper," "also," "provide," "approach," "thus," "main," "obtain," "obtained," "without," "modelling," "modeling," "modeled," "modelled," "use," "result," "results," "resulting," "resulted," "discuss," "discussed," "discussing," "recent," "recently," "give," "gives," "given," "review," "reviewing," "reviews," "require," "required."

We end by stemming the tokens so as to retain only the roots of words in the same family, so as to unify the information contained in related words. For example "education," "educative," and "educated," are all related to education, so we just keep the root "educ" for all of them. The use of these stems relax dimensionality problems and groups all probabilities for families of words into one.

In our sample were initially 13,835 different terms. After this process without loss of generality, we reduce the number of unique terms to 4241 in the corpora with which we build the document term matrix.
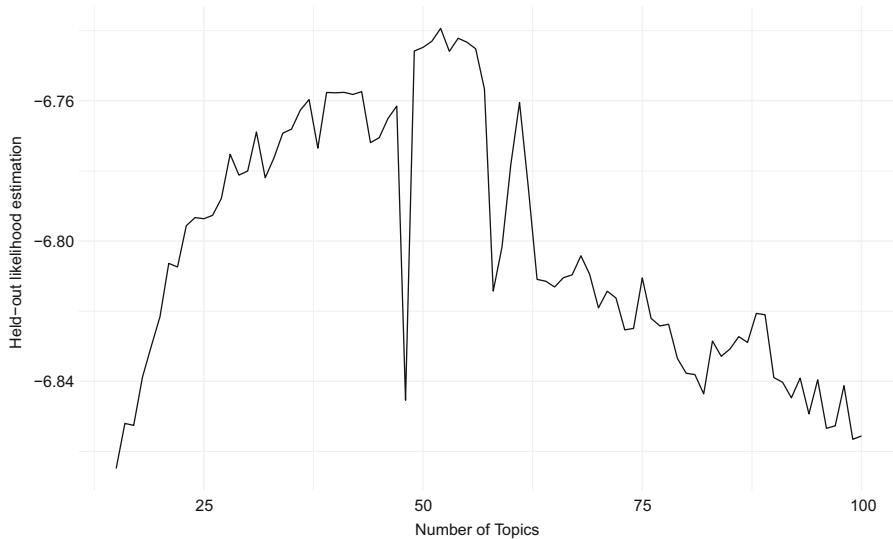
Fig. 21 Held-out likelihood estimation

Appendix C: The Optimal Number of Topics

To run the model involves a choice of hyperparameters as discussed in "Appendix A" above, and one of those parameters is the number of this latent topics existing in our corpus. As this can be interpreted as an arbitrary prior, we run some automatic tests in order to choose this optimal K without human intervention, in order to classify texts in the best possible way. This approach gives us the advantage of automatically selecting the number of topics that better fits data. Arbitrary choosing too few topics means to cluster several topics into a single one. Choosing too many topics means would tend to identify patterns in language rather than topics.

We learn a lot on the different patterns of the data when choosing various alternatives for a fixed number of topics, as we will discuss below. However, our primary selection strategy for automatic selection focuses on the held-out likelihood estimated. Figure 21 reports the log-likelihood of the model evaluated at the estimated parameters on the test set for each K between 15 and 100. The likelihood is maximized between 49 and 54 topics.

Figure 22, in its turn, displays the number of iterations to convergence of the model, which sharply drops at 54 topics and remains at that number of iterations (except for a small spike at 60) beyond 62 topics.

Finally, Fig. 23 reports the semantic coherence which is decreasing and stable after 59 topics. Semantic coherence is maximized when the more frequent words in a given topic co-occur together Mimno et al. (2011). High semantic coherence is reached when in the end there is less topics dominated each by few words. On the other hand, average exclusivity is large when a particular word frequency corresponds to each topic. We follow Roberts et al. (2014) to use the FREX metric for this criteria. As shown in Fig. 24, there are two maximums in 51 and 54 topics.
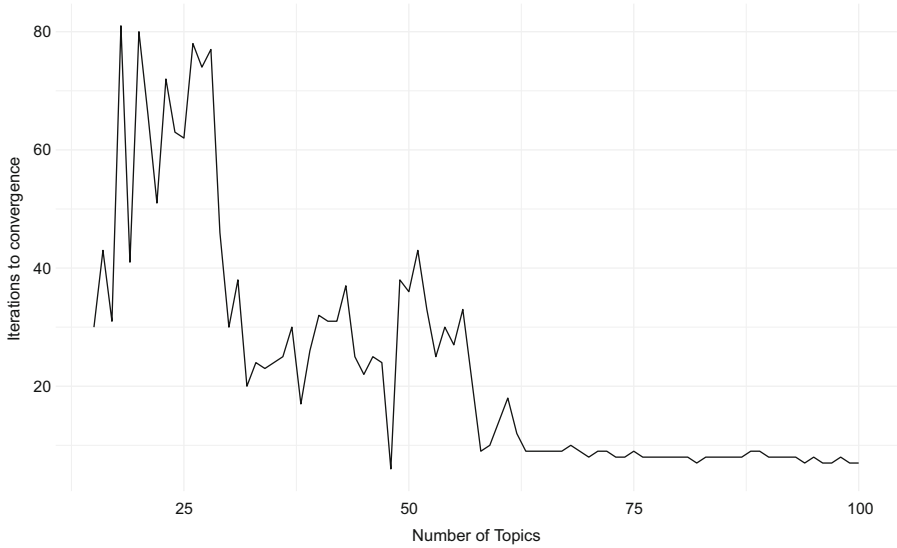
Fig. 22 Number of iterations to convergence of the model
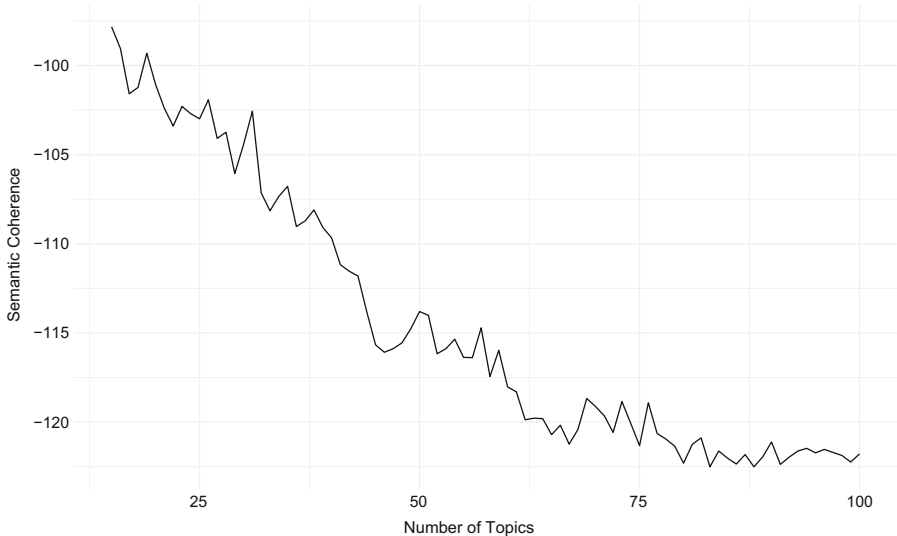
Fig. 23 Semantic Coherence

With our data, we found reasonable to assume that the result is in the neighborhood of 52 topics given the held-likelihood procedure, and given the additional tests, we select the highest number of topics in this neighborhood, corresponding to 54 topics.
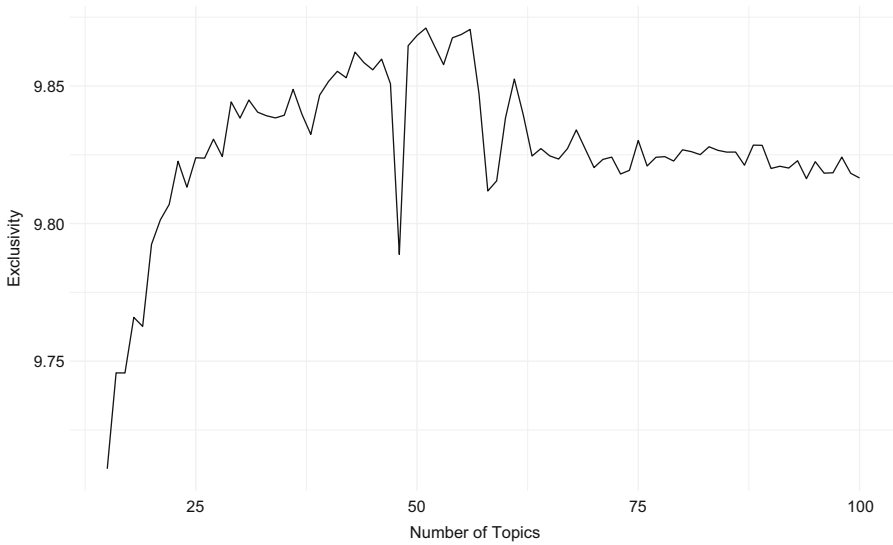
Fig. 24 Exclusivity

Appendix D: The Topics Profile

Given that we have chosen automatically the number of latent topics, it can be helpful to try to disentangle their nature. As an alternative to Figs. 7 and 8, we use simple correspondence analysis to measure the distance between topics. This is a descriptive technique to explore relationships among categorical variables. In our application, we use the matrix of probabilities (the matrix θ_d obtained from STM) for each and every document to belong to any particular built-in topic in order to measure the distance between topics. The rows in this matrix are probabilities that add up to one. The clustering of rows measures the distance between topics (the columns of the matrix). This is the so-called Chi-square distance:

$$\theta_{ij}^{\mathrm{col}} = \sum_{i=1}^{r} \left(p_{ai} - p_{aj} \right)^2 ,$$

where r is the total number of rows, and the measure we compute and represent gives the Euclidean distance between columns i, j(col), for each and every row a (abstract).

Figure 25a depicts the two larger coordinates of the distance matrix computed through classical multidimensional scaling (MDS), so as to obtain the coordinates of the column category. The coordinates are given by the order of largest-to-smallest variance. We find the corpus organized along two dimensions: Dimension 1 can be interpreted as going from Applied to Theory, whereas Dimension 2 goes from, say, Economics to Econometrics. We think this is apparent from casual inspection of Fig. 25a, which involves square distances between $[-4, +4]$.
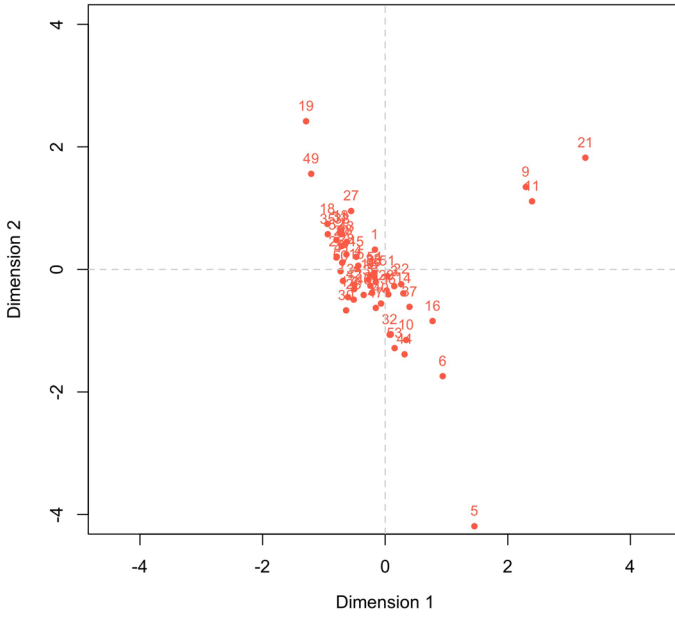
(a) Whole Sample
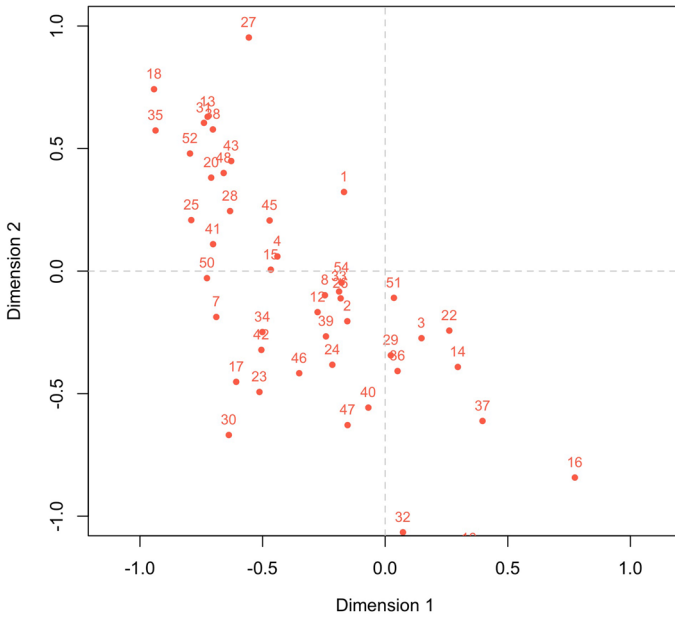
(b) Zoom-in Sample

Fig. 25 Larger coordinates of the distance matrix computed through classical multidimensional scaling (MDS)
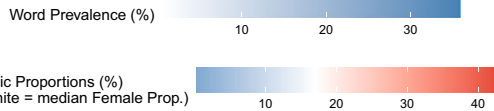
Fig. 26 Latent topics ranked by prevalence in the corpus with $k = 70$. Extended sample with P&P articles

Clearly though, outliers (understood as the topics far away from the origin) are very important in this representation. First, we identify outliers 21, 9, 11, that we have associated with Econometric Theory in the fields of estimation ("estim," "asymptot,"…are the keywords in this case) and testing ("test," "asymptot,"…), together with structural econometrics ("identifi," "instrument,"…), respectively. These actually are among the top 10 more prevalent topics. Moreover, topics 9 and 11 are 2nd and 3rd most prevalent. These outliers are located northeast in the diagram in terms of the language they use.

The second set of outliers are located southeast and are equally far from the center, while not isolated. These topics can be associated with Economic Theory texts. On top of those, we find topic 5, and then not that further away from the center, topic 6, 16 and 10. These are, respectively, auction theory (auction, bid,…), together with game (game, player,…) and information theory (belief, signal,…), as well as mechanism design (mechan, implement,…). These topics are relatively less prevalent in the sample than the Econometric Theory topics above as we discussed in the main text.
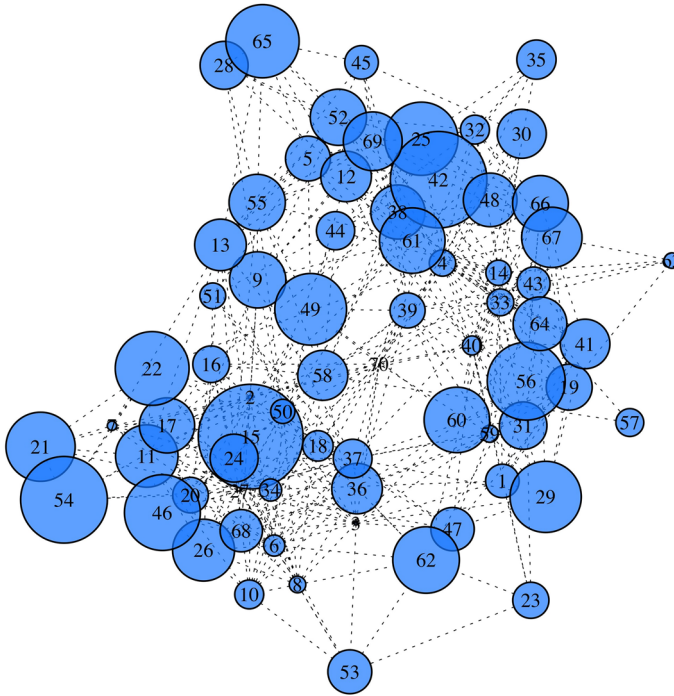
Fig. 27 Connectedness between topics and the fraction documents/abstracts in each topic (θ_d distribution). Extended sample with P&P articles

Finally, there are some outliers at the northwest corner of the diagram. We find here topics that seems to be mostly empirically oriented (applied), and according to our representation, nearly as distant from Econometric than from Economic Theory. These are particularly topics 19 and 49 that we have associated before with Education and Gender issues, and for which female authors' presence is relatively more prevalent.

There is finally a negative correlation between the two coordinates, suggesting that distance values are larger than under the hypothesis of independence between these two key dimensions. This finding would require a treatment that goes beyond the scope in this paper. We leave further analysis of the nature of latent topics in leading economic journals for future research. The interested reader can check the center of the representations at square distances between $[-1, +1]$ in Fig. 25b.

Appendix E: Analysis with the Abstracts of the Papers Proceeding Papers (P&P)

In this section, we extend our original sample with the Papers and Proceedings (P&P) articles published in AER in the especial issue of May during the period 2011–2018.[15]

---

[15] Before 2011, the P&P articles did not have abstract and after 2018 the P&P articles are included in a different journal.
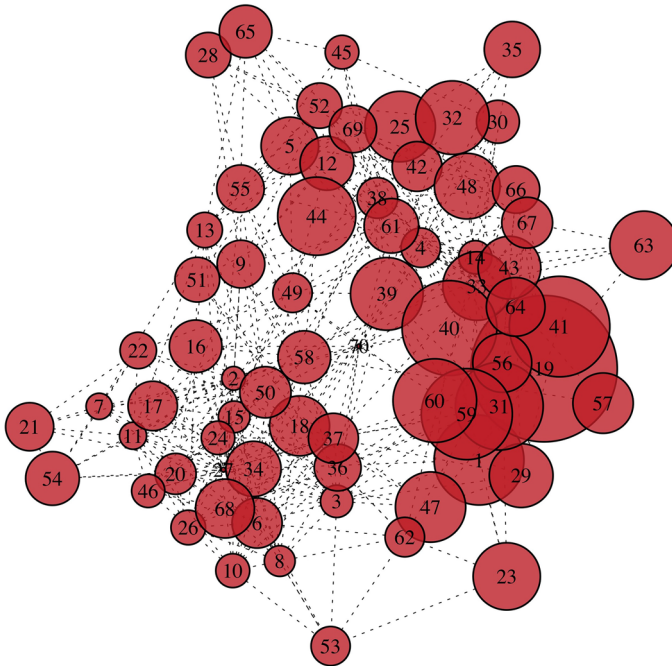
Fig. 28 Connectedness between topics and the female authors documents/abstracts in each topic. Extended sample with P&P articles

These P&P articles are very short (for example, they could be just an extension of a full article submitted to a different journal), and they are selected from the papers presented in the annual January meeting of the American Economic Association's (AEA). Part of the papers are selected directly for the committee's members of the AEA meetings and others are chosen from external proposals of special sessions in AEA meetings.[16] Interestingly for our analysis, papers in P&P are linked to the meeting sessions, and then, they come in groups of three or four papers of a specific topic. Then, the editorial process of this P&P is very different from regular submissions and the set of topics is likely to be more diverse, since some of the special sessions in AEA meeting may be relevant for current policy debate but not necessarily for research. For example, in the issue of May 2020, among others, we can find two sessions and the corresponding articles over "The economics of the health epidemics" or "Is United States deficit policy playing with fire?".

With these additional P&P papers, our sample contains 6428 abstracts/documents, that generates 253,312 tokens and 12,936 unique terms. The number of topics that best fits the these extended sample is 70. The larger number of latent topics can be related to the larger number of unique words and documents, but also to the selection process of P&P described above, sessions unrelated to standard research with a small number of ("seed") papers very related among themselves.

---

[16] For more information about the about the AEA Papers and Proceedings go to: https://www.aeaweb.org/journals/pandp/about-pandp.

Fig. 29 Topic Word Clouds in
the extended sample with P&P
articles
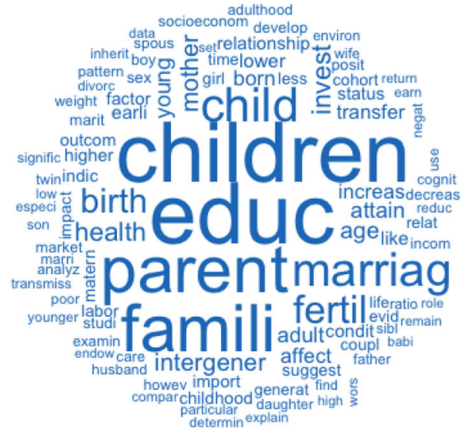
(a) Topic 41

(b) Topic 19

As in the main text, we estimate these 70 latent topics using the STM algorithms.
Figure 26 presents the latent topic ranked by prevalence in the corpus with $k = 70$.

Figure 27 shows the STM output (the estimated latent topics) and also how the
documents are allocated among them.

As in the main text, in Fig. 27 the size of the circle is proportional to the number
of documents in the topic. The most salient feature of Fig. 27 is that in addition to
the larger number of topics, there are some of them with very small size that could
be related to the "seeds" described above, sessions of the AEA meetings, with very
related papers among themselves but quite different to research papers closer to them.

Figure 28 reinforces the evidence of the main message of this paper, male and female
display different pattern when doing research. There is a subset of topics (southeast in
Fig. 28) with a relative high proportion of females, that moreover seems to be closely
connected. On the contrary, there is other set of topic (southwest in Fig. 28) that is
also closely connected and where the presence of females is relatively scarce.

Now, we want to look closer the content of some particular topics. In this larger sample, it is easier to see that the latent topics go beyond standard research fields. In particular, Fig. 29 points out that the latent topics with higher proportions of female authors are topic 41 and topic 19. In the following figure, we can see the distributions over terms that each of this two topic induces are represented as words clouds, where the size of term in the cloud is approximately proportional to its probability in the latent topic distribution β_k. Clearly, topic 41 is related to family economics and topic 19 to gender discrimination.

References

Bagues M, Campa P (2017) Can gender quotas in candidate lists empower women? Evidence from a regression discontinuity design (12149)

Bayer A, Rouse CE (2016) Diversity in the economics profession: a new attack on an old problem. J Econ Perspect 30(4):221–42

Beneito P, Boscá JE, Ferri J, García M (2021) Gender imbalance across subfields in economics: when does it start? J Hum Cap 15(3):469–511

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Boustan L, Langan A (2019) Variation in women's success across PhD programs in economics. J Econ Perspect 33(1):23–42

Buckley C (1985) Implementation of the SMART information retrieval system. Technical Report, USA

Cabrales A, García M, Puch LA (2018) "Gendered Language in the British Press," Mimeo COSME: Gender at 2018 Meetings of the Spanish Economic Association

Card D, DellaVigna S (2013) Nine facts about top journals in economics. J Econ Lit 51(1):144–61

Card D, DellaVigna S, Funk P, Iriberri N (2019) Are referees and editors in economics gender neutral?*. Q J Econ 135(1):269–327

Chari A, Goldsmith-Pinkham P (2017) Gender representation in economics across topics and time: evidence from the NBER summer institute. Working Paper 23953, National Bureau of Economic Research

Chevalier J (2020) The 2020 report of the committee on the status of women in the economics profession

Conde-Ruiz JI, Ganuza J-J, Profeta P (2017) Statistical discrimination and the efficiency of quotas. Fedea Working Papers

Conde-Ruiz JI, Ganuza JJ, Profeta P (2021) Statistical discrimination and committees. Fedea Working Papers (2021-06)

Dolado J, Felgueroso F, Almunia M (2012) Are men and women-economists evenly distributed across research fields? Some new empirical evidence. SERIEs J Span Econ Assoc 3(3):367–393

Gelman A, Meng X, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies. Stat Sinica 733–807

Hansen S, McMahon M, Prat A (2018) Transparency and deliberation within the FOMC: a computational linguistics approach. Q J Econ 133(2):801–870

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

Hengel E (2020) Publishing while female. Are women held to higher standards? Evidence from peer review. Cambridge working papers in economics 1753, Faculty of Economics, University of Cambridge

Hengel E, Moon E (2020) Gender and quality at top economics journals. Working papers 202001, University of Liverpool, Department of Economics

Lundberg S, Stearns J (2019) Women in economics: stalled progress. J Econ Perspect 33(1):3–22

Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) optimizing semantic coherence in topic models, pp 262–272

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B, Rand DG (2014) Structural topic models for open-ended survey responses. Am J Polit Sci 58(4):1064–1082

Roberts ME, Stewart BM, Tingley D (2019) STM: an R package for structural topic models. J Stat Softw Artic 91(2):1–40

Siniscalchi M, Veronesi P (2020) Self-image Bias and lost talent (28308)

Tang C, Ross K, Saxena N, Chen R (2011) What's in a name: a study of names, gender inference, and gender behavior in Facebook, pp 344–356

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.