Gender Distribution across Topics in Top 5 Economics Journals: A Machine Learning Approach

J. Ignacio Conde-Ruiz
Juan-José Ganuza
Manu García
Luis A. Puch

February 2021

Barcelona GSE Working Paper Series

Working Paper nº 1241

Gender Distribution across Topics in Top 5 Economics Journals: A Machine Learning Approach

J.Ignacio Conde-Ruiz[a,c], Juan-José Ganuza[b], Manu García[c] and Luis A. Puch[c]

February, 2021

Abstract

We analyze all the articles published in Top 5 economic journals between 2002 and 2019 in order to find gender differences in their research approach. Using an unsupervised machine learning algorithm (Structural Topic Model) developed by Roberts et al. (2019) we characterize jointly the set of latent topics that best fits our data (the set of abstracts) and how the documents/abstracts are allocated in each latent topic. This latent topics are mixtures over words were each word has a probability of belonging to a topic after controlling by year and journal. This latent 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 uneven distributed along these latent topics by using only data driven methods. The differences about gender research approaches we found in this paper, are "automatically" generated given the research articles, without an arbitrary allocation to particular categories (as JEL codes, or research areas).

Keywords: Machine Learning; Structural Topic Model; Gender; Research fields.

JEL classification: I20, J16

Corresponding author: juanjo.ganuza@upf.edu

[a] Fedea;[b] Universitat Pompeu Fabra and Barcelona GSE,[c] Universidad Complutense de Madrid and ICAE;

We thank Antonio Cabrales, Pedro Delicado and Nagore Iriberri for helpful comments. José Ignacio Conde-Ruiz and, Manu García and Luis Puch, respectively, acknowledge the Spanish Ministry of Science and Innovation for financial support through projects PID2019-105499GB-I00 and PID2019-107161GB-C32. Juan-José Ganuza gratefully acknowledges the nancial support from the Spanish Agencia Estatal de Investigacin, through the Severo Ochoa Programme for Centres of Excellence in RD (CEX2019-000915-S) and the Spanish Ministry of Education and Science through project ECO2017-89240-P.

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 (2019) (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 very often requires to publish in Top 5 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 United States Economics departments and conclude that Top 5 publications are a very powerful explanatory variable of the promotion to Tenure. Publishing in a Top 5 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 on these journals is also determining in some way 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 articles published 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 declining number of published papers in Top

[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)).

5. Card et al. (2019) analyze two of the Top 5 journals (the *QJE* and *REStud*) and include *Journal of European Economic Association* and *Review of Economics and Statistics*, and report that the current proportion of accepted papers is 3%. Is this Top 5 entry barrier harder for women? The answer provided by Card et al. (2019) to this question is ambiguous. They analyze whether or not females are discriminated in the evaluation of their submissions to their set of leading journals. On the one hand, authors do not find any gender biases in the referees editorial recommendations and editors decisions are gender-neutral conditional on the referee advises. On the other hand, however, they find that conditional on referee recommendations, female authored papers end up accumulating more citations in later years.[2]. A potential explanation for second result is that journals hold female-authored papers to higher standards, but it also could be related to some "horizontal" features or characteristics of female-authored papers that lead to more citations but not to higher acceptance rates in the editorial process. As Card et al. (2019) control by research areas (JEL codes) their results may be linked to more subtle "horizontal" differences, for example, that in the same research line, male choose a more theoretical approach and females a more applied perspective. The methodology we are using in this paper allow us to identify these subtle gender "horizontal" research differences.

It is important to study the distribution of men and women across research topics in order to understand their performance gap in their publishing and tenure achievements. In fact, 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. Women are particularly underrepresented in macro, finance and economic theory, and more prominent in labor

[2]In the same line, Hengel and Moon (2020) analyze publications in Top 5 and they also find that females published articles are more cited.

and other applied micro-economic fields. Beneito et al. (2018) find similar results using data from the annual AEA meetings from 2010-2016, Lundberg and Stearns (2019) focus on PhDs dissertation in Economics from 1991-2017, representing almost all major PhD-granting departments in the United States. Using JEL code for identifying the research area, 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.

In this paper, we want to contribute to this literature in two directions. We focus on exploring gender "horizontal" distribution across research topics in the Top 5 publications. More importantly, we do so, using a new methodological approach based on Machine Learning techniques. We collect all articles published in Top 5 journals for the period 2002-2019. We obtained 5,311 articles, and we take for each article authors' names, year of publication, journal and the abstract. With this information, we provide a very accurate picture of the performance of men and women in publishing in these leading journals.

Second, we use a Machine Learning algorithm to classify our abstracts' database into latent topics. In particular, from the universe of algorithms for topic modelling we implement and develop the Structural Topic Model (STM) developed by Roberts et al. (2019) because it allows 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 an stochastic model that generates "latent topics' and allocate the documents to them. The main advantage of this unsupervised machine learning approach is that "latent topics" are mixtures over words were each word has a probability of belonging to these topics, and these topics can capture, 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. Our main result is that female are unevenly distributed across these latent

topics. We show also that although the proportion of females is slightly increasing among the population of Top 5 authors, these "horizontal" differences persists. We have computed the empirical density distribution of latent topics by gender conditioning of having published in the Top 5, and we show some striking differences between male and female. We want to emphasize the importance of these results, not only because latent topics may capture more subtle "horizontal" differences, but also because the differences about gender we estimated are "automatically" generated given the documents, without an arbitrary allocation to particular categories (as JEL codes, or research areas).

Finally we reduce the number of topics in the algorithm trying to capture the mixtures of words that determine the research areas. There is a trade off when choosing the number of topics. On one hand, a high number of topics usually fits better the data. On the other hand, a lower number of latent topics facilitates the semantic interpretation of them. In our setting, it makes latent topics more alike to standard research fields. Consistently with our previous finding, we also find an uneven distribution of topic/research fields by gender, very much in line with the existing literature cited above.

There are several channels for which the gender differences in the choice of research topic that we have identified in this paper 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 these "horizontal" gender differences plus an initial uneven distribution of gender researchers may generate an unintentional discrimination trap linked with the functioning of academic organizations (journals, departments, etc.). Conde-Ruiz et al. (2017, 2021) analyzes 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 abstract topic 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 generates 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) focuses specifically on academic labor market and points out a similar unintentional discrimination trap linked to the so-called "*self image bias*". Research evaluators are biased towards 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 evaluators, 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 paper is organized as follows: the next section presents the raw data and the descriptive analysis of the patterns of publication in Top 5 journals; section three presents the Structural Topic Model; section four study the gender differences in the latent estimated topics; section five extends the model to analyze topics as research fields. Section six concludes and in the Appendix we explore several extensions and provide more 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 Top 5 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.

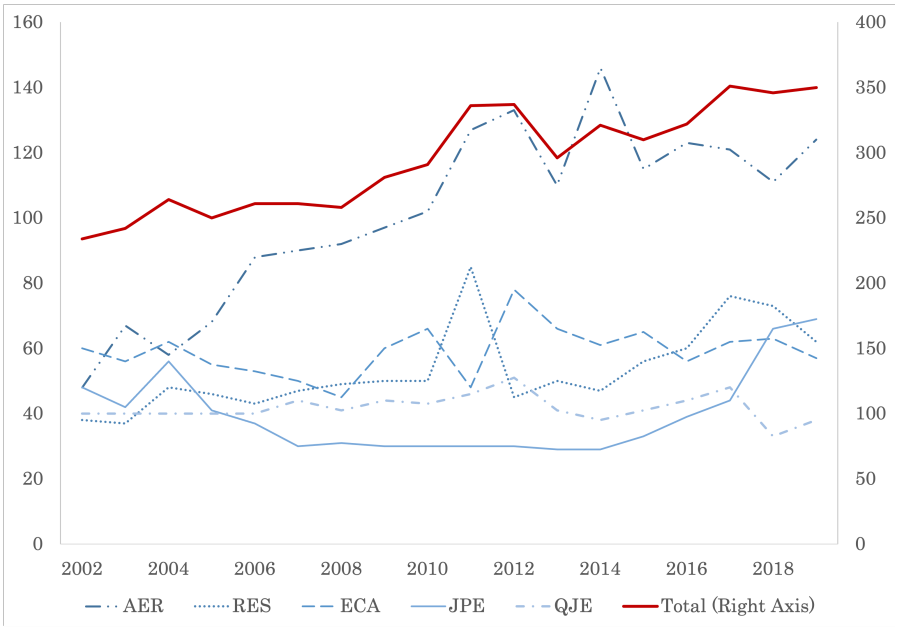


Figure 1 : Number of Articles Published per Year in Top 5[3]

We have 5,311 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 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 their editorial process are different[5]. We want to compare this descriptive information with Card and DellaVigna (2013) who analyze all the articles published in the Top 5 from 1970 to 2012. They obtain several interesting facts, among them that the total number of articles

[3]Publications exclude notes (without abstract), comments, announcements, and Papers and Proceedings (P&P).

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

[5]In the Appendix E we add P&P articles to our data and we replicate the analysis for this extended data base.

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 top 5 publications, up from 25%in the 1970s. In our more updated sample, as it is shown in the Figure 1, we find that this trend have stabilize 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 authors case (or even the four authors case) exhibits the opposite pattern. The two authors 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.

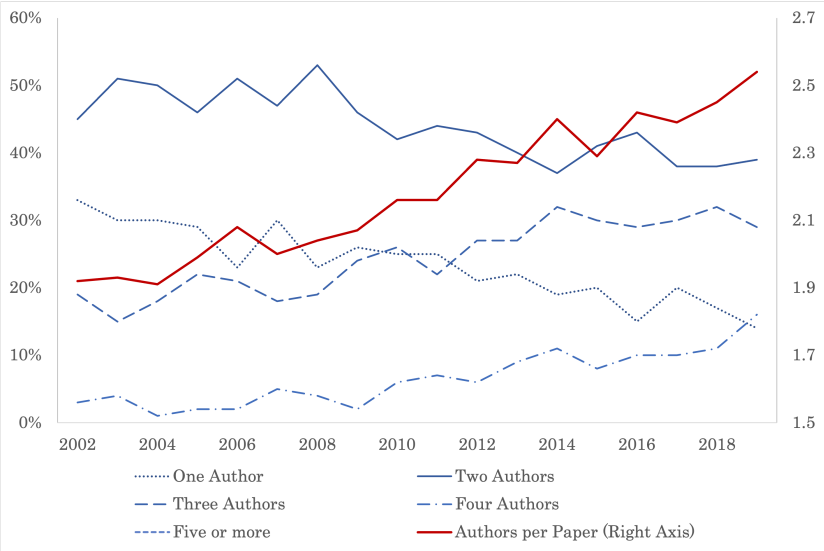


Figure 2 : Number of Authors of Published Papers on Top 5.

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 name. For this purpose, we rely on three different databases: the first-names database published by the U.S. 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 transforms the original 5,311 articles to a total sample of 11,721 (with implied 9,840 articles-men authors, and 1,881 articles-women authors). Except otherwise indicated all measures below are computed over this augmented articles-authors sample.

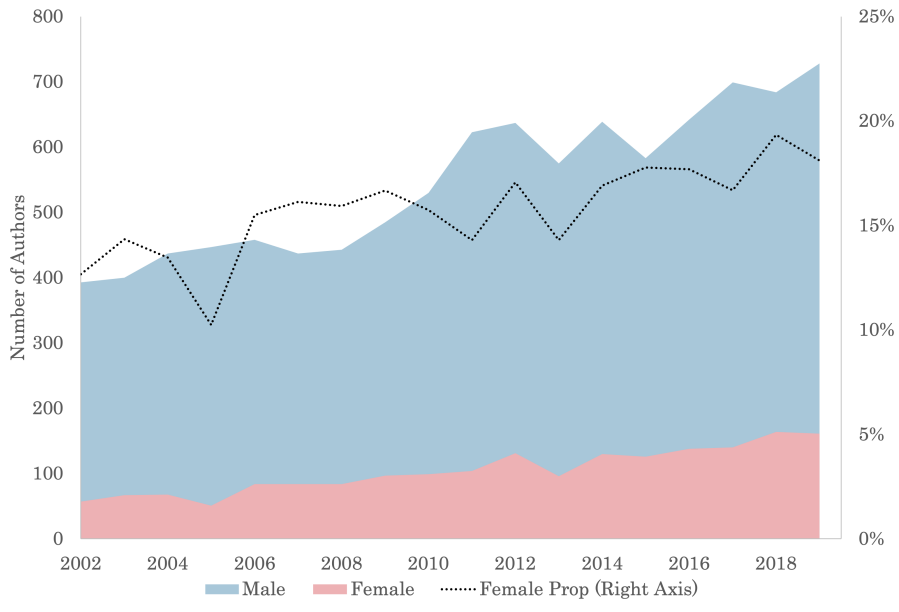
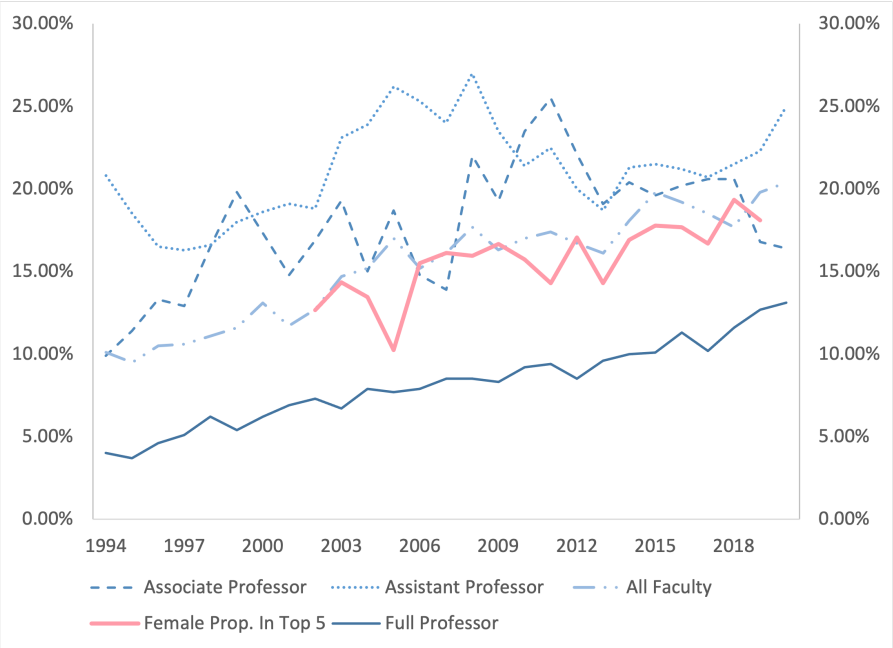


Figure 3 : Number of article-author observations by gender, and the share of female articles.

Figure 3 depicts the share of female authors (right axis), which has been steadily increasing (with fluctuations) at a rate of 6.2% per year, (compared to mens 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 there have been an important improvement in the last decades, women are clearly under-represented in Top-5 publications. These data are consistent with the other 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 in the United States 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% of share of women in the faculty of the top 20 Schools in the United States in 2020. In line with Heckman and Moktan (2020), the rate of increasing of female coauthors in Top 5 seems to be very similar to the rate of increasing of female full Professors in these Universities.



Source: CSWEP Report, 2020 and own elaboration.

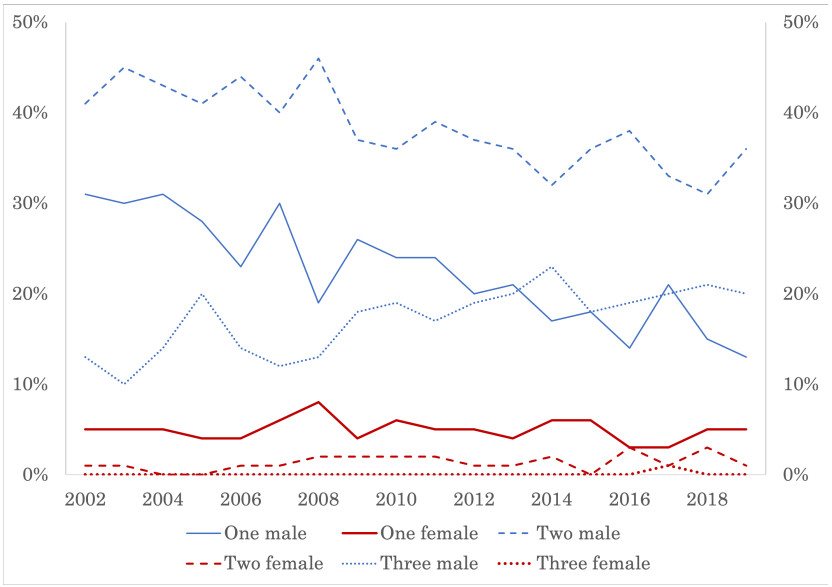
Figure 4 : The Pipeline for Top 20 Departments: Percent and Numbers of Faculty and Students who are Women.

Finally, we move to analyze the pattern of co-authorships for male and females. We have split the description of the data in two figures, one for single gender groups and another for mixed teams. Figure 5a shows the corresponding co-authorships pattern when the set of co-authors are single gender groups.

The more salient feature of these data are that, while the share of male solo authors has been declining from 30% of total, to slightly above 10%, the share of female solo articles has been stable over the entire sample, at a share close to 5%. We want also to point out that despite the slow declining, two males is the most common co-authors team.

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

[h!]



(a) Percentage of articles coauthored by single gender teams.

(b) Percentage of articles coauthored mixed gender teams.

Figure 5 : Co-authorships patterns in Top 5 journals.

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 to the participation of female authors in articles with more male authors.

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]. In this context, by unsupervised we denote the absence of human intervention in order to identify the latent topics behind the abstracts of the articles published in Top 5 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 over a particular topic there are words that are used more often than others. Then our objective is to provide a low-dimensional representation (topics) of a high dimensional object (abstracts) while keeping as much as possible its informational content.

The baseline for topic modelling 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 5,311 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,

[6]For an excellent non technical introduction to machine learning, see Hansen et al. (2017).

[7]For technical description of the LDA algorithm, see the original article of Blei et al. (2003) and also Hansen et al. (2017) that is the first paper that uses the LDA algorithm in the economic literature.

after converting to lower case and remove from the original raw text common stop-words[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 4,241 of unique words.

The second step is to represent our text data in a document-term matrix of D rows (5,311 abstracts) and V columns (4,182 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 our document-term-matrix in two 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 speak about several topics. Then, θ_d^k would mean the weight of topic k in document d. Then our a 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).

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 that maximizes

[8]In particular, we remove the stop-words from the SMART list, developed at Cornell University in the 1960

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

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

this total likelihood. We are going to skip here further details on the algorithms we use, and we refer the interested reader to the 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[10]. 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 allow us to go beyond pre-labelled 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 reach field. For this reason, in the

[10]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

next section we will analyze our stochastic model with $K = 54$ topics, while in Section 5, we will be focusing on estimating our stochastic model with $K = 15$ topics. In addition to these two exercises, in the appendix we extend our original sample for including the abstracts of 1,117 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.[11]. We estimate probabilities for each document to belong to this set of built-in latent topics using the Structural Topic Model. The STM output is summarized by the latent topics displayed in Figure 6 that shows the key words associated to 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 issues. In the same line, the second more prevalent topic in the distribution, topic 9, may be associated to Econometric Theory ("*consist*" and "*asymptot*" stems are there). However, this is not the goal of the analysis as we have indicated above. Latent topics may be related to something beyond research fields, as methodology or style of writing.

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

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

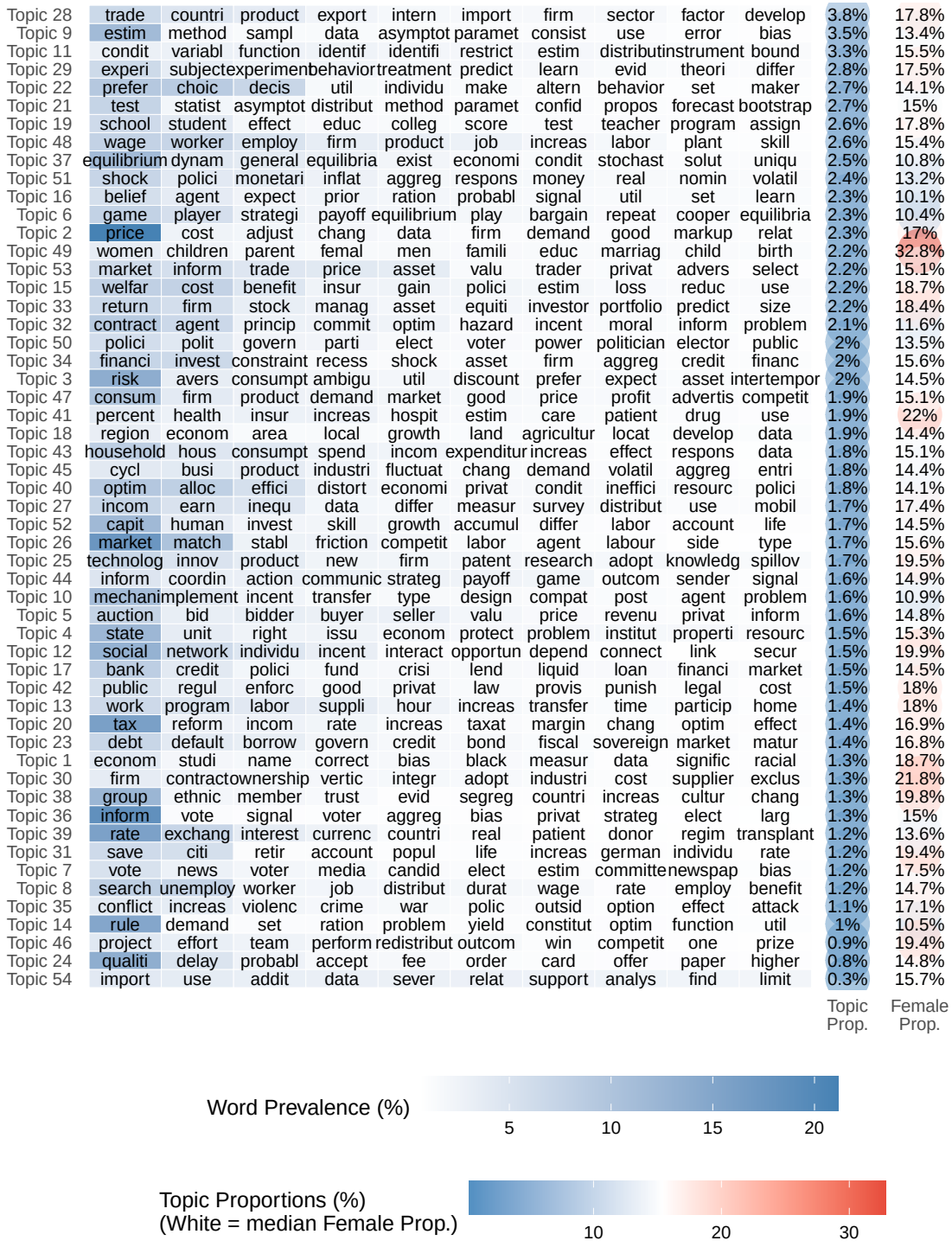


Figure 6 : Optimal K Topics Ranked by Prevalence in the corpus.

topics with different probability weights. Following this approach, the next figure shows that latent estimated topics in a way that also illustrates the number of documents in each topic.

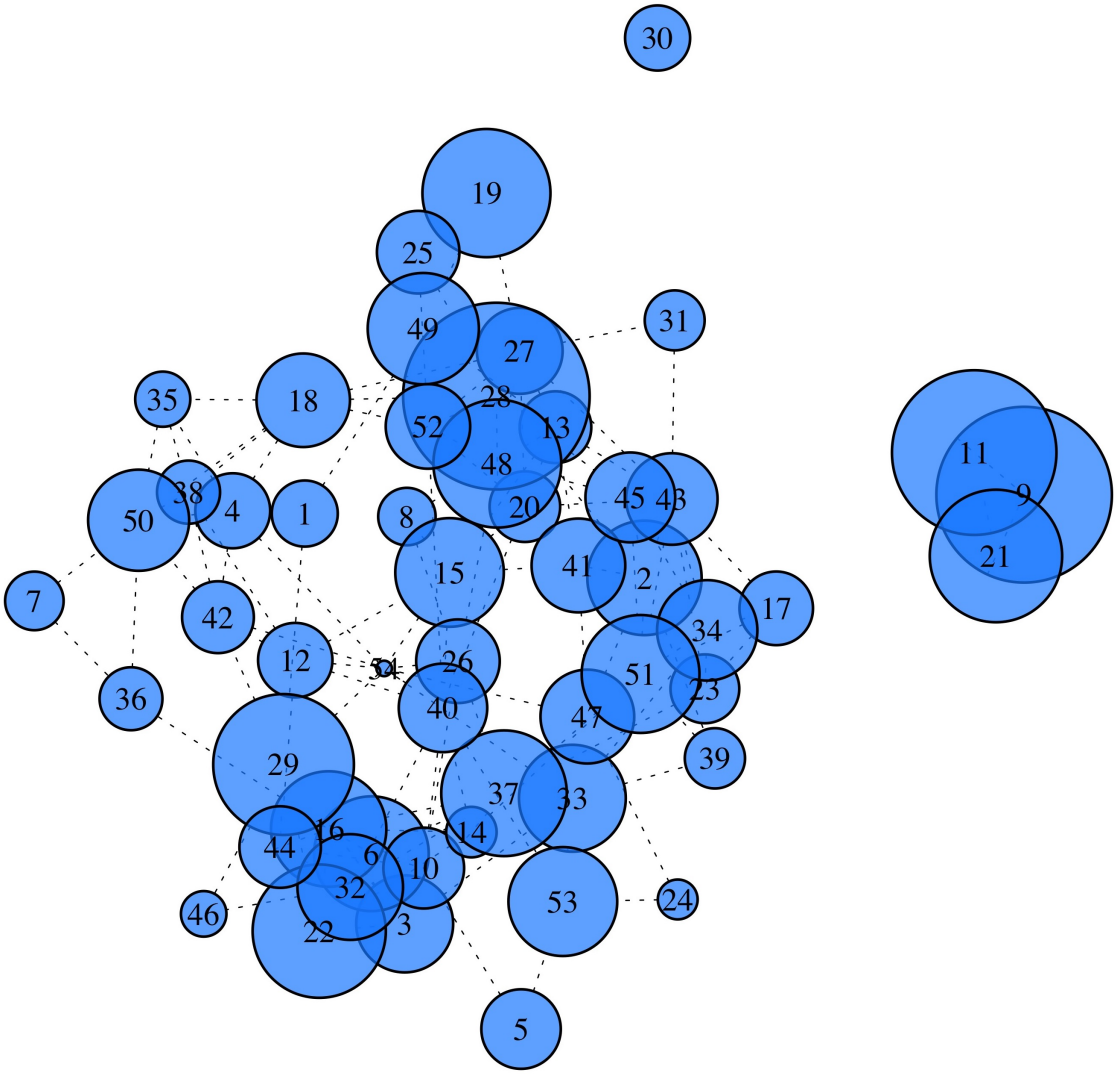


Figure 7 : Connectedness between topics and the fraction documents/abstracts in each topic (θ_d distribution).

Notice that in Figure 7 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 Figure 6). As we cannot make a mapping of our 54 topics to particular fields of research, it is difficult to interpret the information of Figure 7 regarding the size of the topics. For example, topics 11, 9 and 21, in Figure 7 seem very 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 7 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 Figure 6 and the description of the latent topics in Figure 7, 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 Figure 7 we can also identify some clusters of topics, for example (West in Figure 7) 51,34, 23, 2, etc are topics related to Macro-Finance); (East in the Figure 7) 50 is a central node of a set of topics related to Political Economy and Institutions), (South-West in Figure 7) 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 Figure 7). In Appendix C 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.

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 8 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 Figure 6).

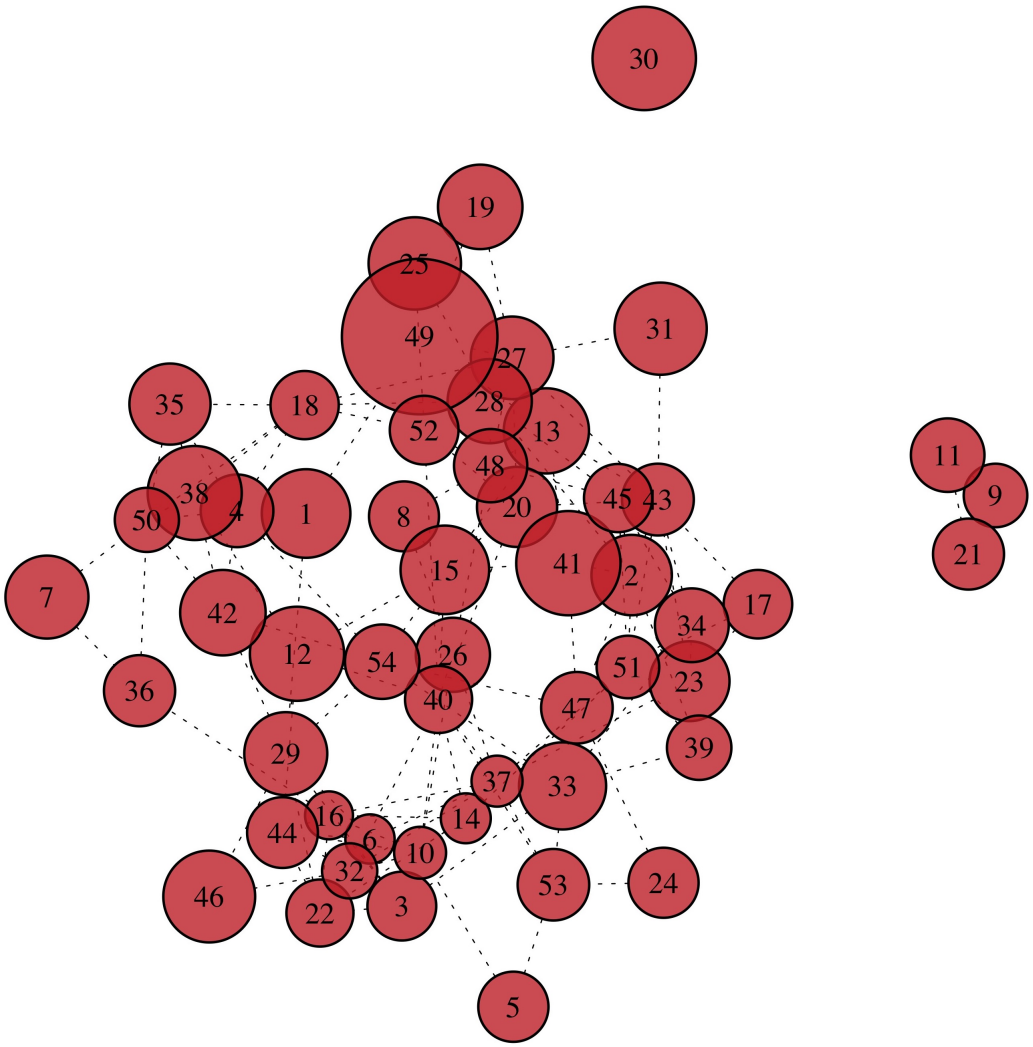


Figure 8 : Connectedness between topics and the female authors documents/abstracts in each topic.

Figure 8 provides interesting evidence of the main message of this paper, male and female displays different patterns when doing research. Independently of the grade of underrepresentation 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 the figure 8), specially topic 49, with a relative high proportion of females, that moreover seems to be closely connected (around

the terminology used related to applied economics fields). On the contrary, there is other set of topics (for example South-West in Figure 8) that are also closely connected and where the present of females is scarce (around terms common to economic theory research questions).

As we said above, it is very difficult to describe the precise semantic meaning of the latent topics when we are working with $K = 54$. We want, however, to look closer the latent topics where females are more or less prevalent. In particular, Figure 9 shows that the latent topic with the highest proportion of female authors is topic 49. On the contrary topic 16 turns out to be the topic with the lowest proportion of females. Figure 9 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.

(a) Topic 49 (highest proportion of female authors).

(b) Topic 16 (lowest proportion of female authors).

Figure 9 : Topic Word Clouds: Topic 49 vs Topic 16

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

Figure 10, shows the mean and the standard deviation across time of the presence of

women authors by topic and provides a clear evidence over the "horizontal" gender differences in research. We show that for some latent topics the proportion of females is larger than the average (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 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). As we show, in Figure 3, the proportion of female authors has been increasing in the last two decades from around 13% on 2002 to 19% on 2019. It is possible that latent topics that have received a large part of these new female coauthors have higher average of females and as the same time a larger dispersion over time.

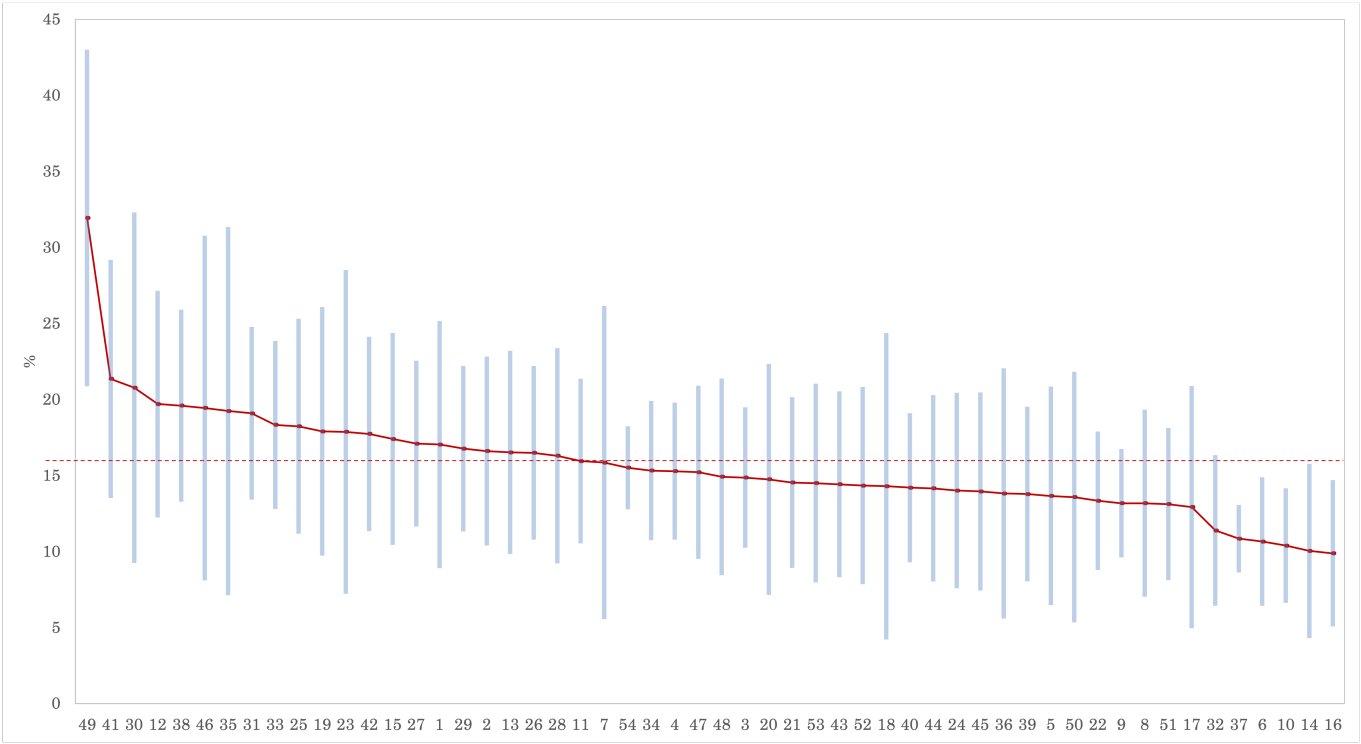


Figure 10 : On the presence of women, by topic: mean and one standard deviation across time.

Using the data of Figure 10, we compute the average dispersion (the variance) of the proportion of females authors with respect to the mean in the period 2002-2019. Figure 11 displays this dispersion by year, as a first approximation to know if the horizontal differences between male and female, decrease or increase along time.

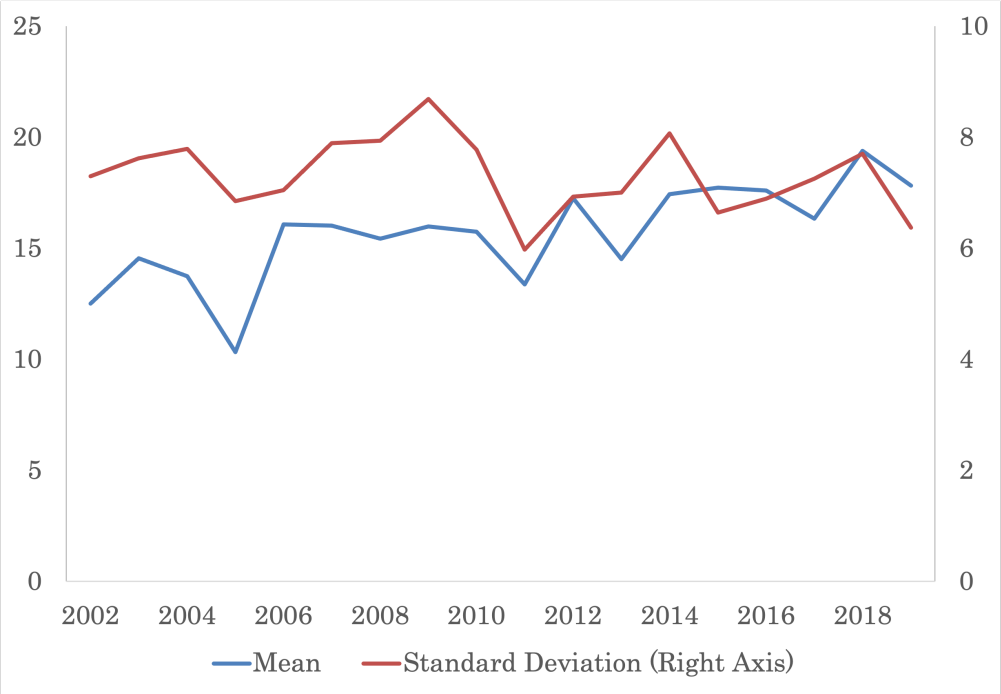


Figure 11 : On the presence of women, by year: mean and one standard deviation across time.

Figure 11 shows that there is not a clear trend in the dispersion of the proportion of females by topic. Therefore, Figure 11 illustrates that the gender gap is slowly shrinking regarding the proportion of females authors in published papers in Top 5, but we do not observe that the distribution of females across topics is going to converge to a more even distribution.

The dispersion across topics is a signal of gender "horizontal" differences in research. However, 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 empirical density distributions across topics between males and females, conditional of having published an article in Top 5. Once a female have published a paper, this density function give us the probability that this paper belong to one of the 54 topics. We rank the topics according to probability of being chosen by a male author.

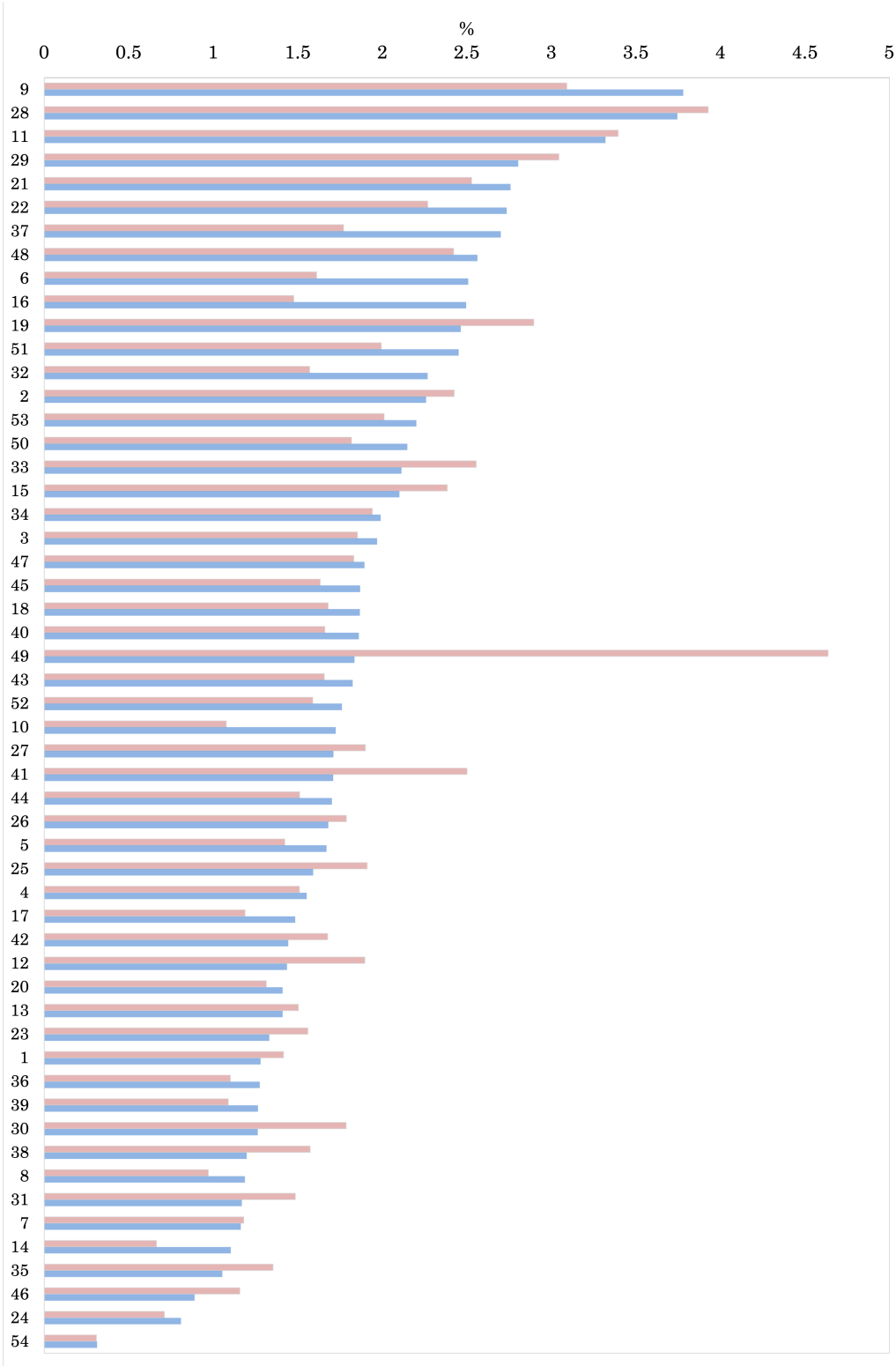


Figure 12 : Empirical density distributions across topics between males and females (conditional of having published an article in Top 5).

Figure 12 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 empirical density 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.

Finally, we confirm this intuition with a complementary Figure 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 Figure 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.

Figure 13 shows that there are topics for which the propensity of publishing papers by females is higher than males, and the opposite. 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.

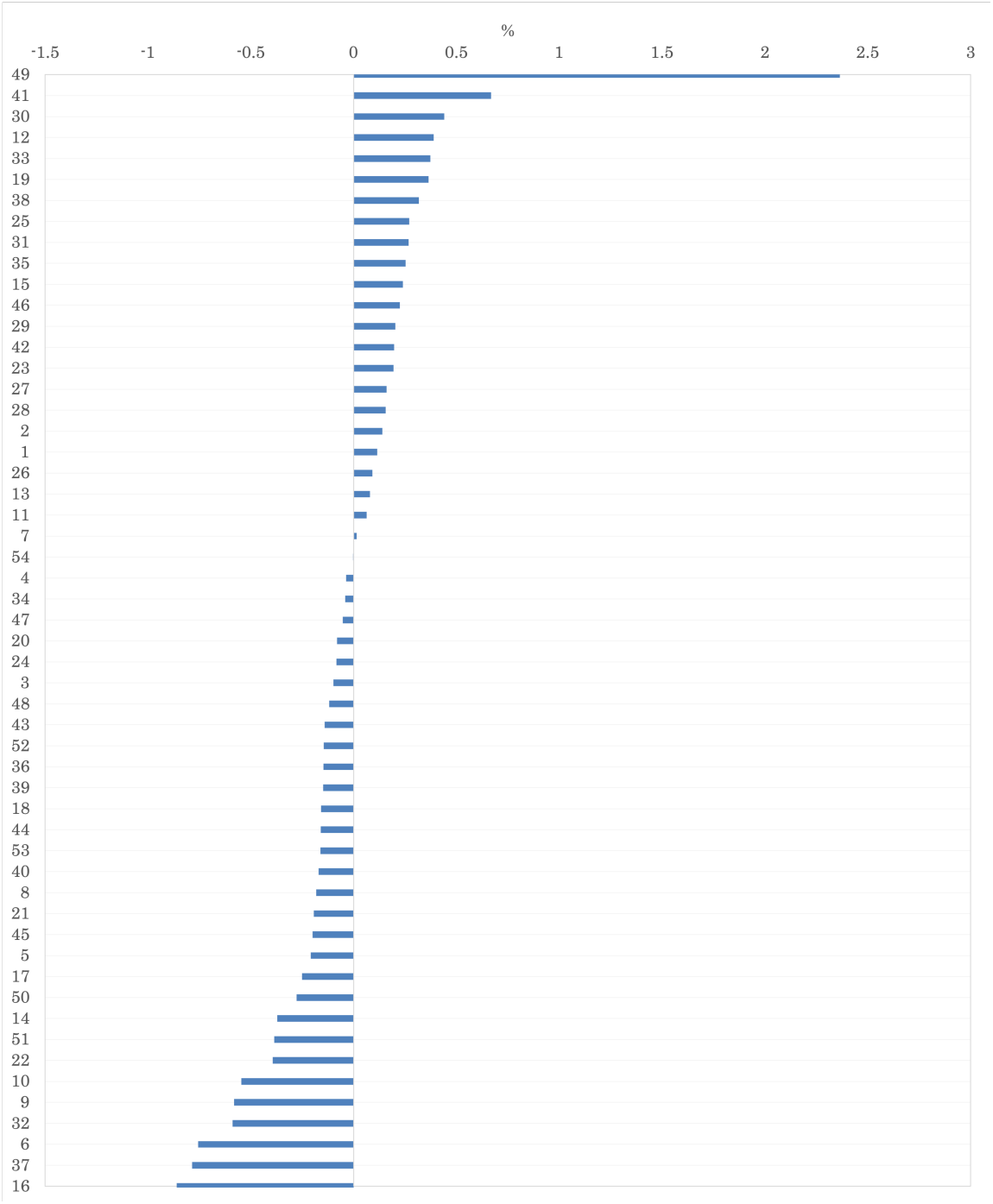


Figure 13 : Relative propensity of publishing papers by females over topics.

5 Topics as Research Fields

In this section we estimate the stochastic model with a lower number of topics, with two objectives. On 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 Top 5 has increased over time. On the other hand, a low number of topics will allow us to frame our results with previous literature that have used a small number of categories linked to JEL codes and research areas in top departments. After estimating the model for a range of $K \in 10, ..., 20$, we have found that $K = 15$ is a number of topics for which the estimated model performs better in terms of fitting with the data, and in terms of the semantic content of the latent topics at the same time. The model with $K = 15$ latent topics is summarized in Figure 14.

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. Following our previous analysis, next Figure 15a plots the latent topics showing the relative semantic distance between topics as well as their weight in terms of the fraction documents/abstracts that they content.

If we compare Figure 7 (with $K = 54$) and Figure 15a (with $k = 15$), they have 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 as Labor and Public Economics, are closely connected.

Figure 15b (as Figure 8 with $K = 54$) provides evidence of the "horizontal" differences between male and female in doing research. This results go in line with the previous literature as in Dolado et al. (2012), Chari and Goldsmith-Pinkham (2017), Beneito et al. (2018) and Lundberg and Stearns (2019) that point out that females are unevenly distributed across fields. We coincide with previous literature that females are over-represented in Applied-Micro fields, specially Health-Gender and Experiments and Education and underrepresented in Economic Theory fields and Macro-Monetary and Finance.

For example, Dolado et al. (2012) use the classification of women by research areas

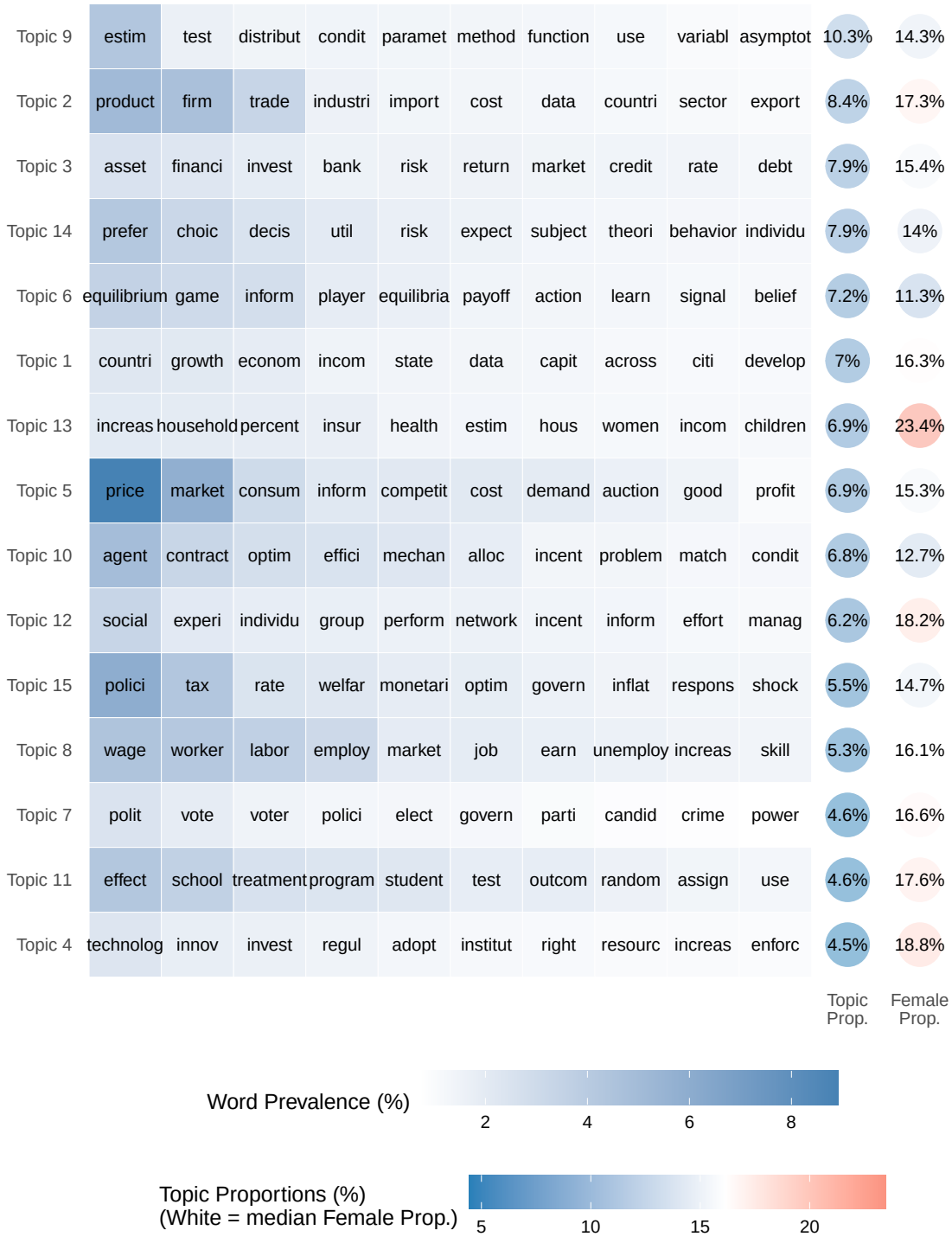
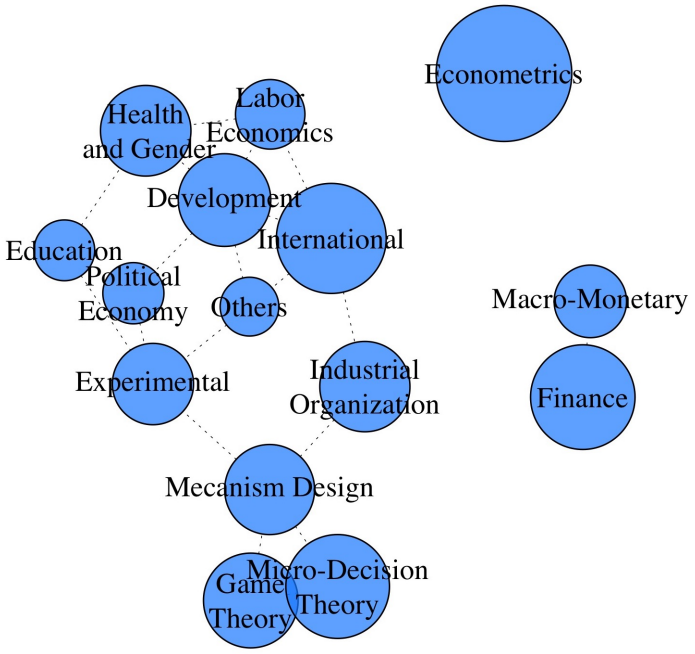
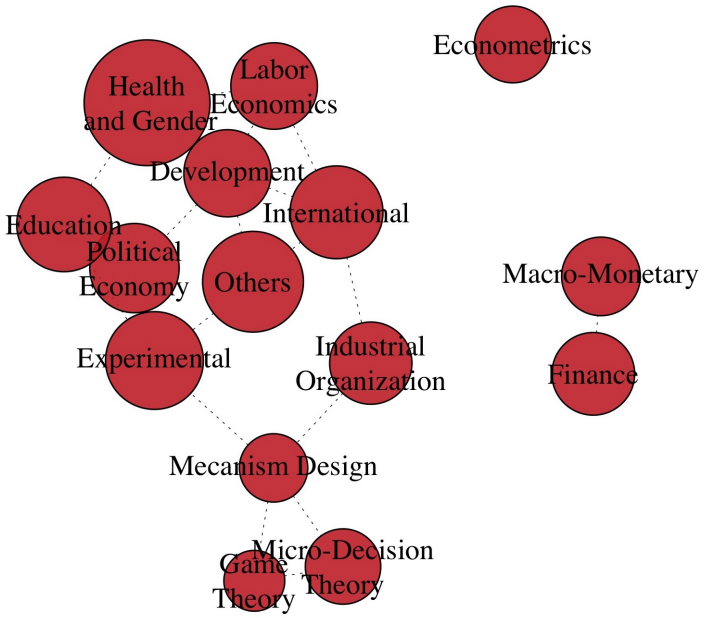


Figure 14 : Latent topics ranked by prevalence in the corpus with $k = 15$.

(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



(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.

Figure 15 : Connectedness for $K = 15$

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 by topics seems to be slightly less disperse than those identified by previous literature. This can be due that our methodology is more "continuous" than allocating females to fixed categories, as far as the probabilistic model allocates females' articles to latent topics with statistical weights.

Figure 16 analyzes together the evolution of the prevalence of the topics and the proportion of females authors.

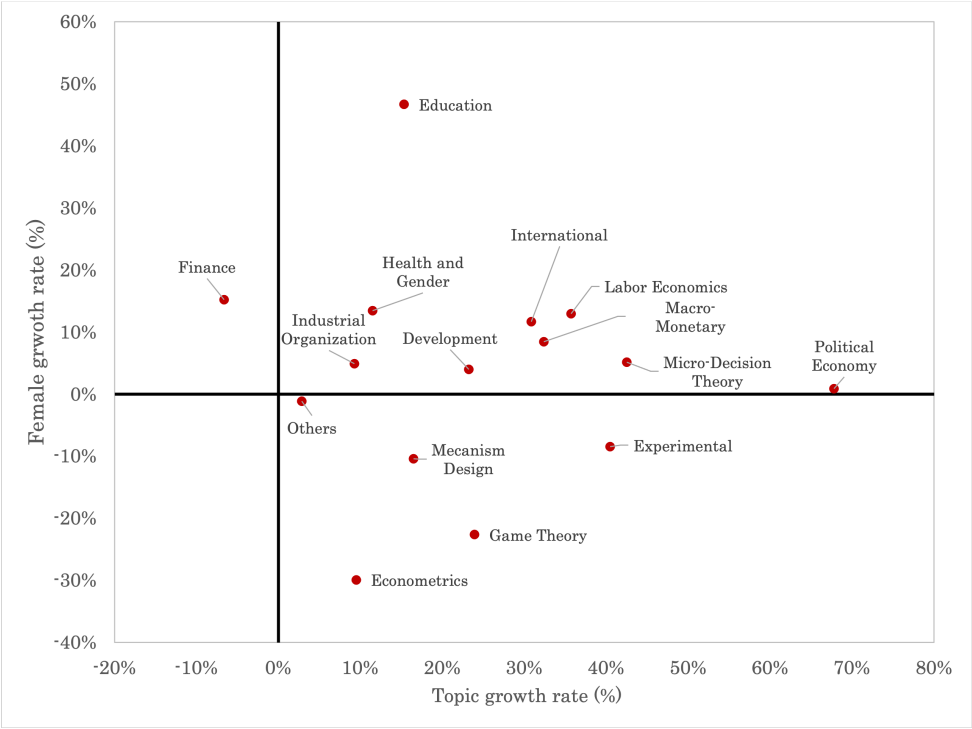


Figure 16 : Growth rates of prevalence and female proportion by topics.

For building Figure 16, we have computed the growth rate of topics' prevalences and topics' female proportions using the latest seven years (2013-2019) and the first seven years (2002-2008). 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 not a clear relationship between the growth rate of topic prevalence and the increase in female representation. 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 Top 5 market should be relatively large. New entrants should be more likely to work in "hot" topics rather 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.

6 Conclusions

Using a new data base composed by the abstracts of all articles published in Top 5 journals in Economics for the period (2002-2019) and by using unsupervised machine learning techniques, we have shown that there are persistent and significant differences in the way males and females approaches econ research. Using the Structural Topic Model we have identified 54 latent topics and shown that the distribution of female authors among them is uneven. These findings are important for several reasons, because: i) Top 5 publications are key for research careers and also for determining the path of the 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 field of research. iii) Finally, recent theoretical results 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 three 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. Secondly, we want to analyze more formally the observed gender differences in the empirical conditional density distributions on the published papers in Top 5. 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.

References

Bagues, Manuel and Pamela Campa, "Can Gender Quotas in Candidate Lists Empower Women? Evidence from a Regression Discontinuity Design," 2017, (12149).

Bayer, Amanda and Cecilia Elena Rouse, "Diversity in the Economics Profession: A New Attack on an Old Problem," *Journal of Economic Perspectives*, November 2016, *30* (4), 221–42.

Beneito, P., J. E. Boscá, J. Ferri, and M. García, "Women across Subfields in Economics: Relative Performance and Beliefs," *Fedea Working Papers*, June 2018.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, March 2003, *3* (null), 9931022.

Boustan, Leah and Andrew Langan, "Variation in Women's Success across PhD Programs in Economics," *Journal of Economic Perspectives*, February 2019, *33* (1), 23–42.

Buckley, Chris, "Implementation of the SMART Information Retrieval System," Technical Report, USA 1985.

Cabrales, A., M. García, and L. A. Puch, "Gendered Language in the British Press," *Mimeo COSME: Gender at 2018 Meetings of the Spanish Economic Association.*, 2018.

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

_ , _ , Patricia Funk, and Nagore Iriberri, "Are Referees and Editors in Economics Gender Neutral?*," *The Quarterly Journal of Economics*, 11 2019, *135* (1), 269–327.

Chari, Anusha and Paul Goldsmith-Pinkham, "Gender Representation in Economics Across Topics and Time: Evidence from the NBER Summer Institute," Working Paper 23953, National Bureau of Economic Research October 2017.

Chevalier, Judy, "The 2020 Report of the Committee on the Status of Women in the Economics Profession," 2020.

Conde-Ruiz, J. Ignacio, Juan-José Ganuza, and Paola Profeta, "Statistical Discrimination and the Efficiency of Quotas," *Fedea Working Papers*, 2017.

_ , Juan José Ganuza, and Paola Profeta, "Statistical Discrimination and Committees," *Fedea Working Papers*, February 2021, (2021-06).

Dolado, Juan, Florentino Felgueroso, and Miguel Almunia, "Are men and women-economists evenly distributed across research fields? Some new empirical evidence," *SERIEs: Journal of the Spanish Economic Association*, September 2012, *3* (3), 367–393.

Hansen, Stephen, Michael McMahon, and Andrea Prat, "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach," *The Quarterly Journal of Economics*, 10 2017, *133* (2), 801–870.

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

Hengel, Erin and Eunyoung Moon, "Gender and quality at top economics journals," Working Papers 202001, University of Liverpool, Department of Economics February 2020.

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

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum, "Optimizing Semantic Coherence in Topic Models," 2011, p. 262272.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley, "stm: An R Package for Structural Topic Models," *Journal of Statistical Software, Articles*, 2019, *91* (2), 1–40.

_ , _ , _ , **Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand**, "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, 2014, *58* (4), 1064–1082.

Siniscalchi, Marciano and Pietro Veronesi, "Self-image Bias and Lost Talent," December 2020, (28308).

Tang, Cong, Keith Ross, Nitesh Saxena, and Ruichuan Chen, "What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook," 2011, pp. 344–356.

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 $\text{Mult}_K(\theta_d, 1)$ that indicates the position n associated to a topic.

3. Sample the indicator $w_{d,n}$ from $\text{Mult}_V(B_{z_{d,n}}, 1)$, 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 θ_d, that is,

$$\theta_d \sim \text{LogisticNormal}_{K-1}(\mathbf{\Gamma}' \mathbf{x}'_d, \Sigma), \quad \mathbf{\Gamma} = [\gamma_1|...|\gamma_K]$$

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 θ_d consists of:

   - sampling the probability $\mathbf{z}_{d,n}$ that a word is in a topic: $\mathbf{z}_{d,n} \sim MN_K(\mathbf{\Theta}_d)$, with K outcomes
   - conditional on topic, choose a word from $\beta_{z_{d,n}}$, that is $\mathbf{w_{d,n}} \sim MN_V(\beta_{z_{d,n}}), over \mathbf{B} = [\beta_1|...|\beta_K]$ 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 don't 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, the "common' words such as "and", "for", "in", etc. We removed the stop words 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. This is the list of removed words: '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'.

Finally, we end by stemming the tokens so as to retain only the roots of words in the same family, in order to unify the information contained in related words. For example "education", "educative", and "educated", are all related with education, so we just keep the root "educ" for all of them. For our purposes, we want to know the relation of all these

words with other words. 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 4,241 in the corpora with which we build the document term matrix.

Appendix C The optimal number of topics

To run the model involves a choice of hyperparameters as discussed in Apendix 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.

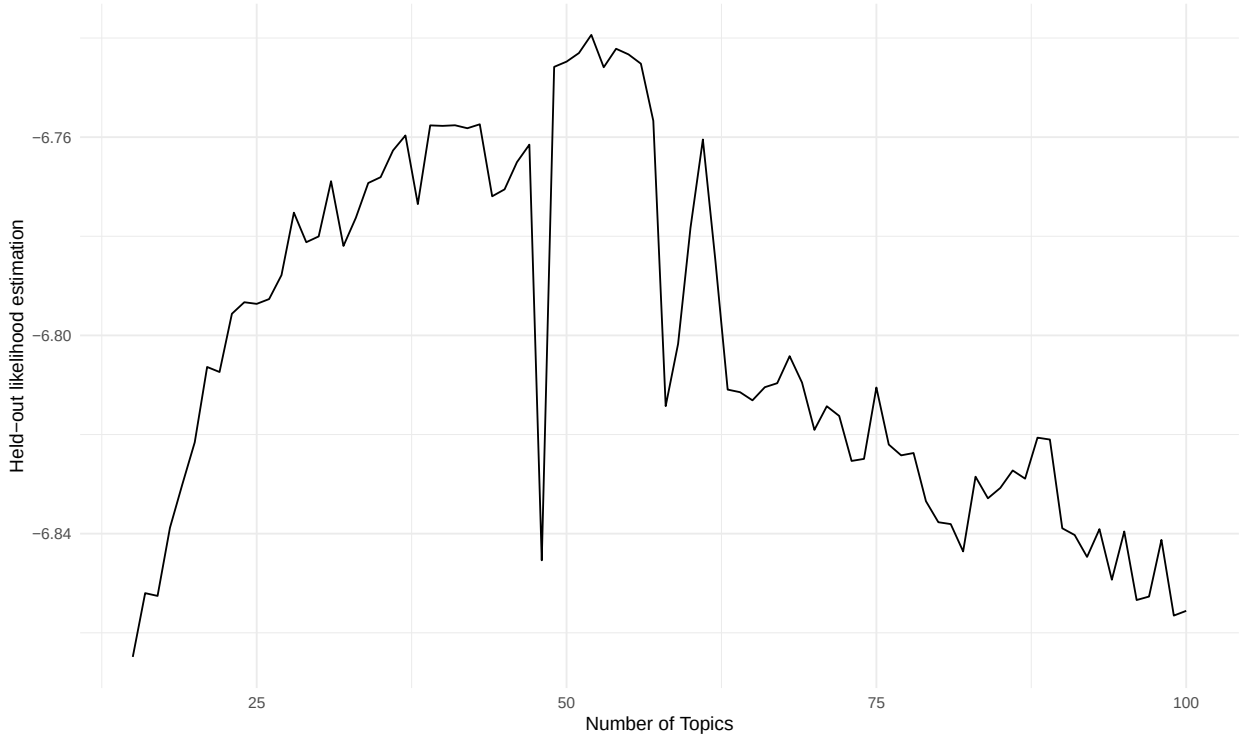


Figure A. 1 : Held-out likelihood estimation

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 A.1 reports the log-likelihood of the model evaluated at the estimated parameters on the test set for each

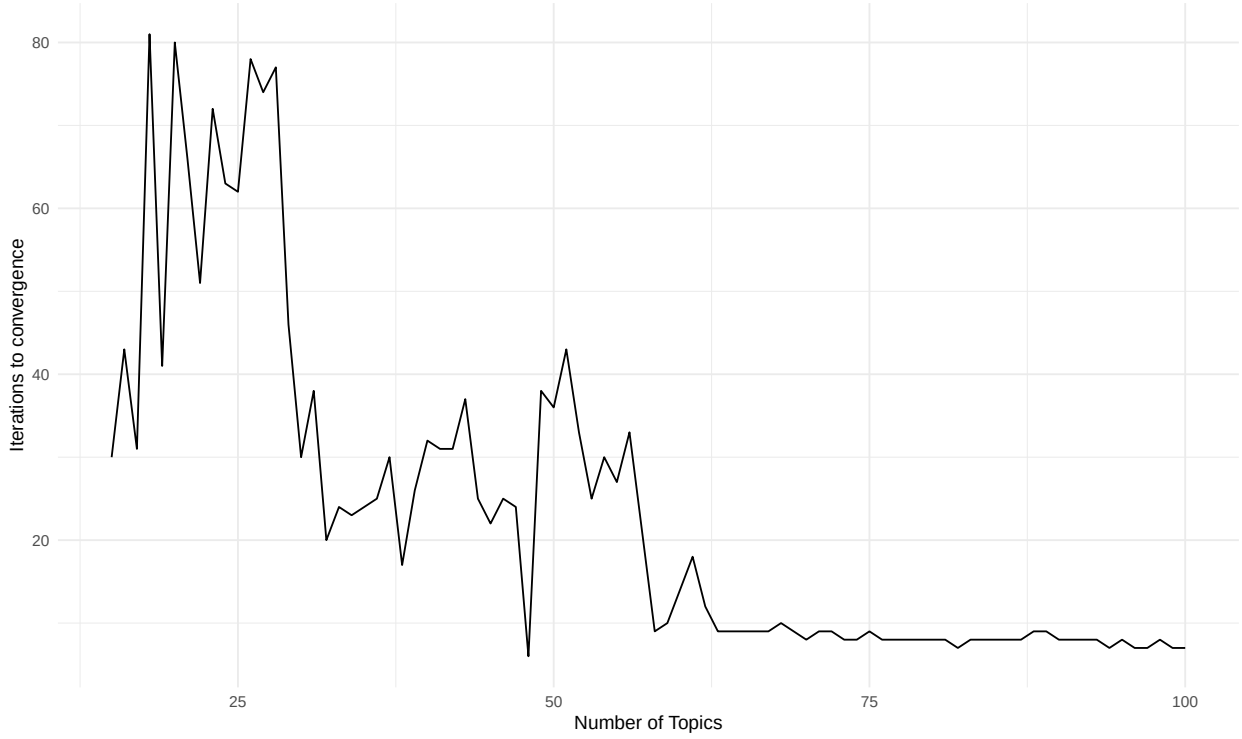


Figure A. 2 : Number of iterations to convergence of the model

K between 15 and 100. The likelihood is maximized between 49 and 54 topics.

Figure A.2, 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, Figure A.3 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 showed in Figure A.4 there are two maximums in 51 and 54 topics.

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.

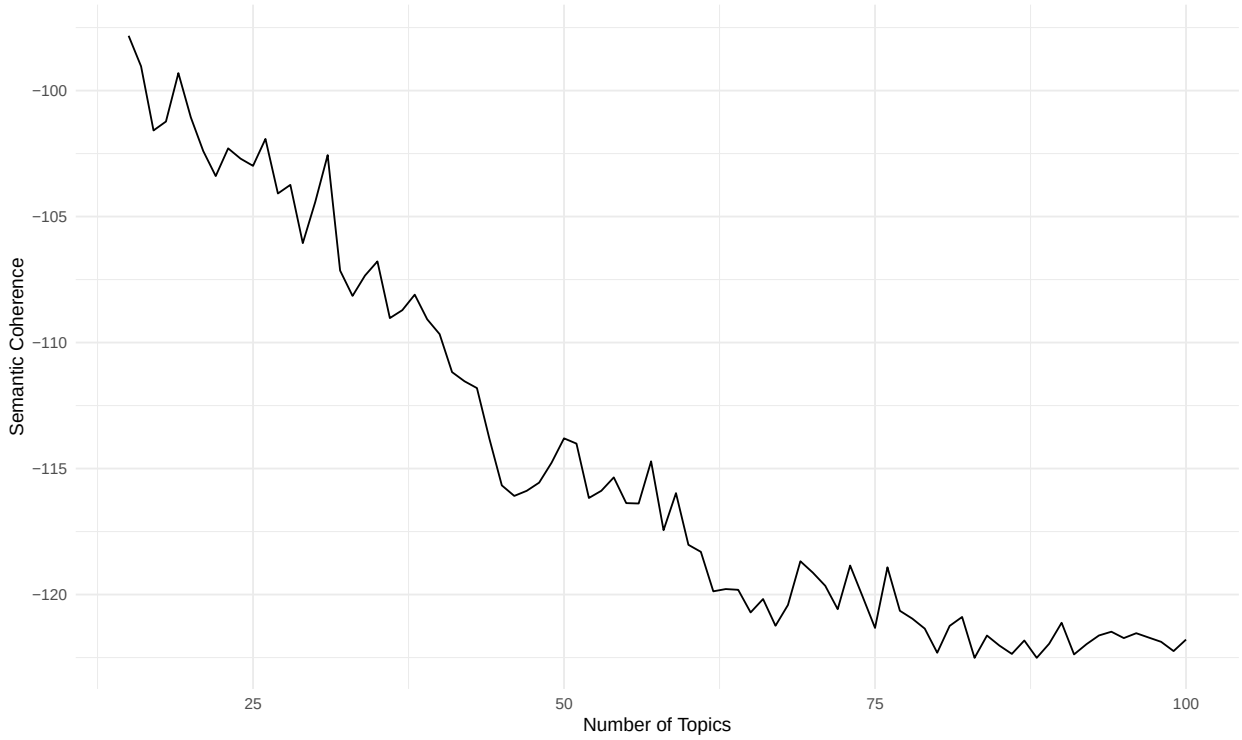


Figure A. 3 : Semantic Coherence

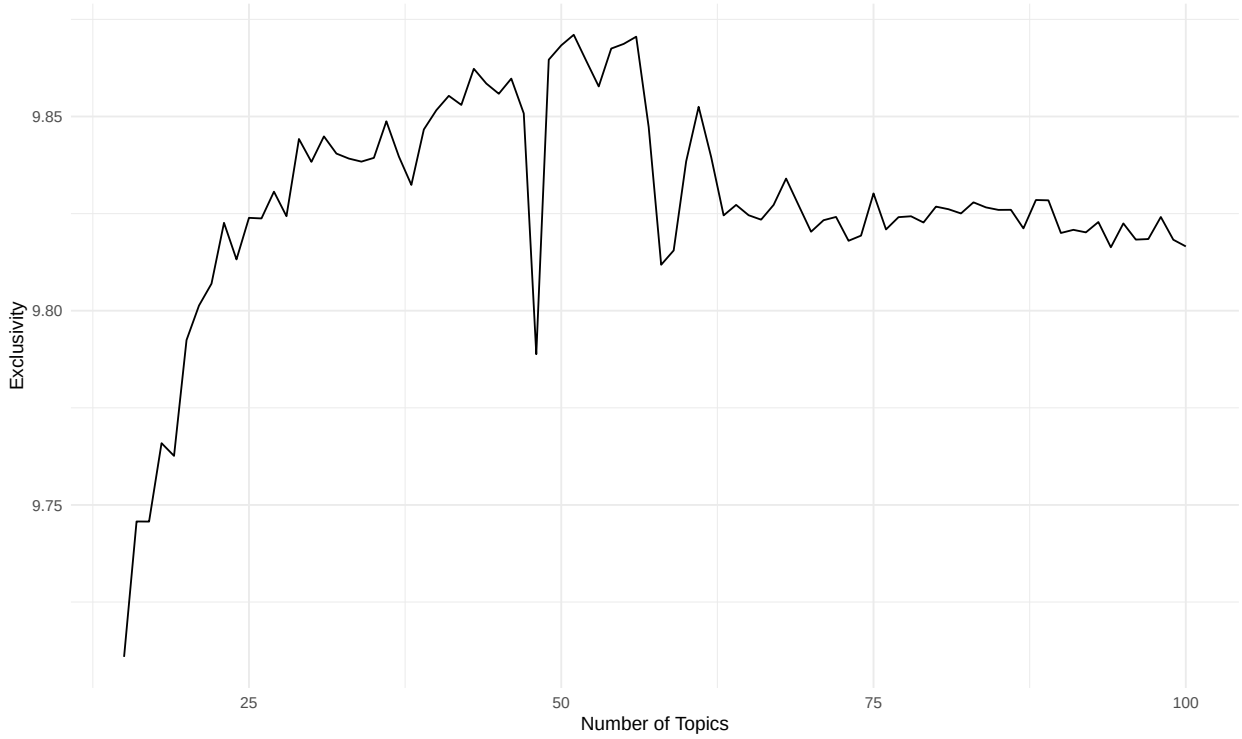


Figure A. 4 : Exclusivity

Appendix D The topics profile

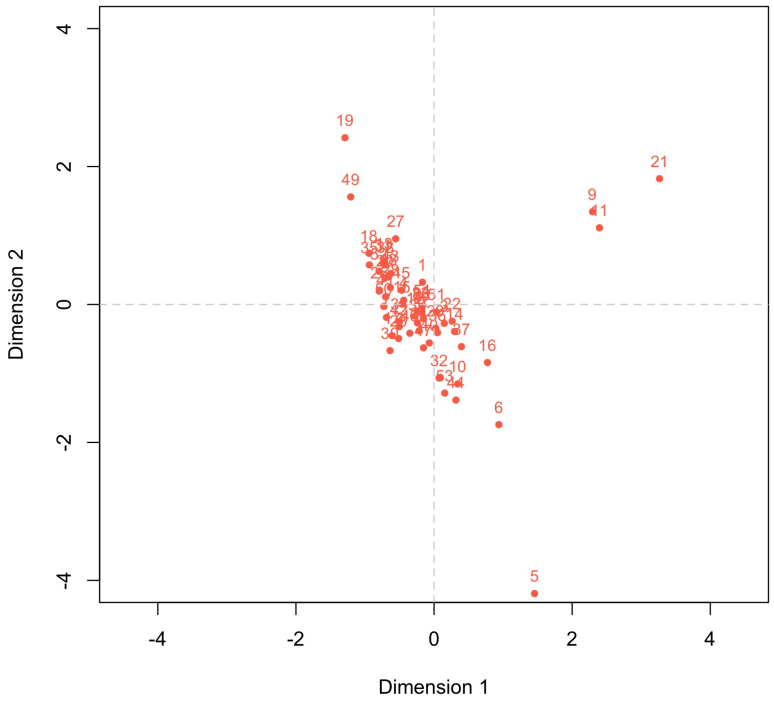
Given that we have chosen automatically the number of latent topics, it can be helpful to try to disentangle their nature. As an alternative to Figures 7 and 8, we use Simple Correspondence Analysis to measure the distance between topics. This is a descriptive technique to explore relationships among 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:

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

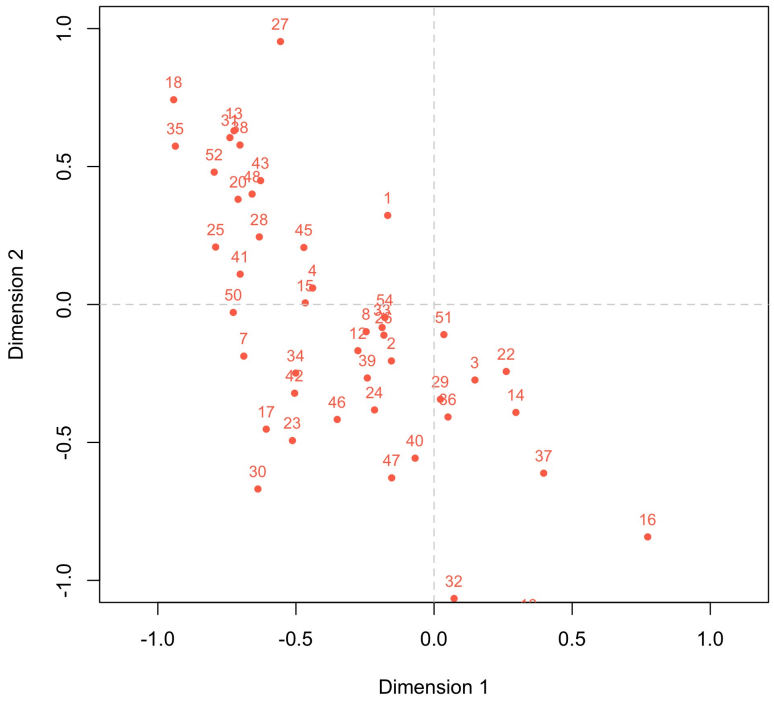
where r is the total number of rows, and the measure we compute and represent gives the euclidean distance between columns i, j (col), for each and every row a (abstract).

Figure A.5a 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 Figure A.5a,which involves square distances between $[-4, +4]$.

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 to 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 are among the top 10 more prevalent topics. Moreover, topics 9 and 11 are 2nd and 3rd most prevalent. These outliers



(a) Whole Sample



(b) Zoom-in Sample

Figure A. 5 : Larger coordinates of the distance matrix computed through Classical Multidimensional Scaling (MDS)

are located North East in the diagram in terms of the language they use.

The second set of outliers are located South East and are equally far from the center, while not isolated. These topics can be associated to 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.

Finally, there are some outliers at the North West 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 Figure A.5b.

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[12]. 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[13]. Interestingly for our analysis, papers in P&P are linked to the meeting sessions, and then, they come in groups of 3 or 4 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 6,428 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. As in the main text, we estimate these 70 latent topics using the STM algorithms. Figure A.6 show the STM output (the estimated latent topics) and also how the documents are allocated among them.

[12]Before 2011 the P&P articles did not have abstract and after 2018 the P&P articles are included in a different journal.

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

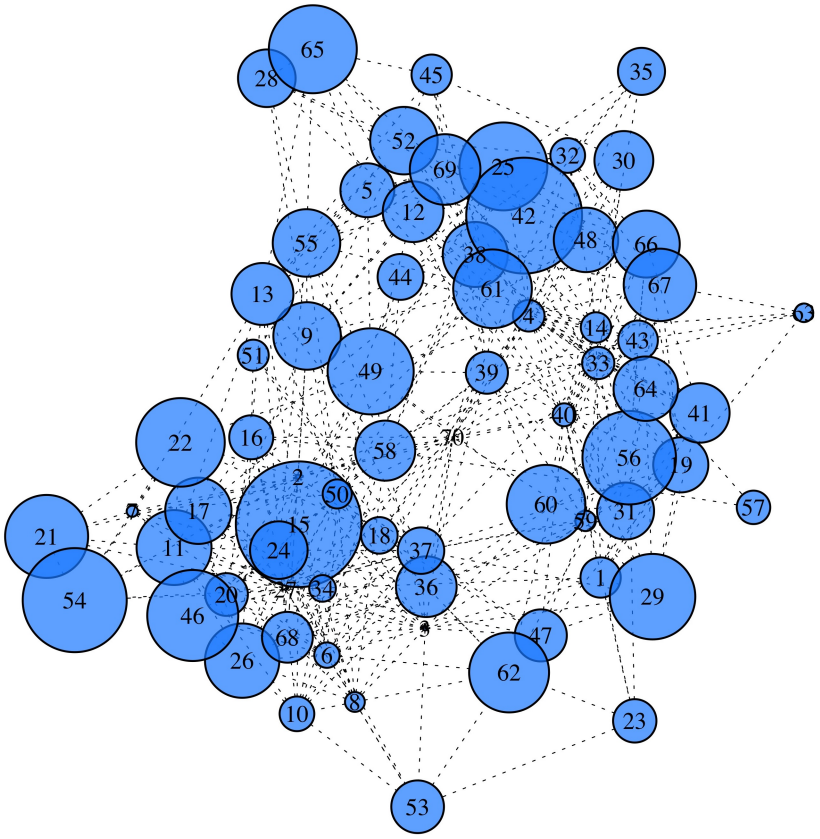


Figure A. 6 : Connectedness between topics and the fraction documents/abstracts in each topic (θ_d distribution). Extended sample with P&P articles.

As in the main text, in the Figure A.7 the size of the circle is proportional to the number of documents in the topic. The most salient feature of the Figure A.7 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 A.7 reinforce the evidence of the main message of this paper, male and female display different pattern when doing research. There is a subset of topics (South-East in the figure A.7) with a relative high proportion of females, that moreover seems to be closely connected. On the contrary, there is other set of topic (South-West in the Figure A.7) that is is also closely connected and where the present of females is relatively scarce.

Now, we want to look closer the content of some particular topics. In this larger sample,

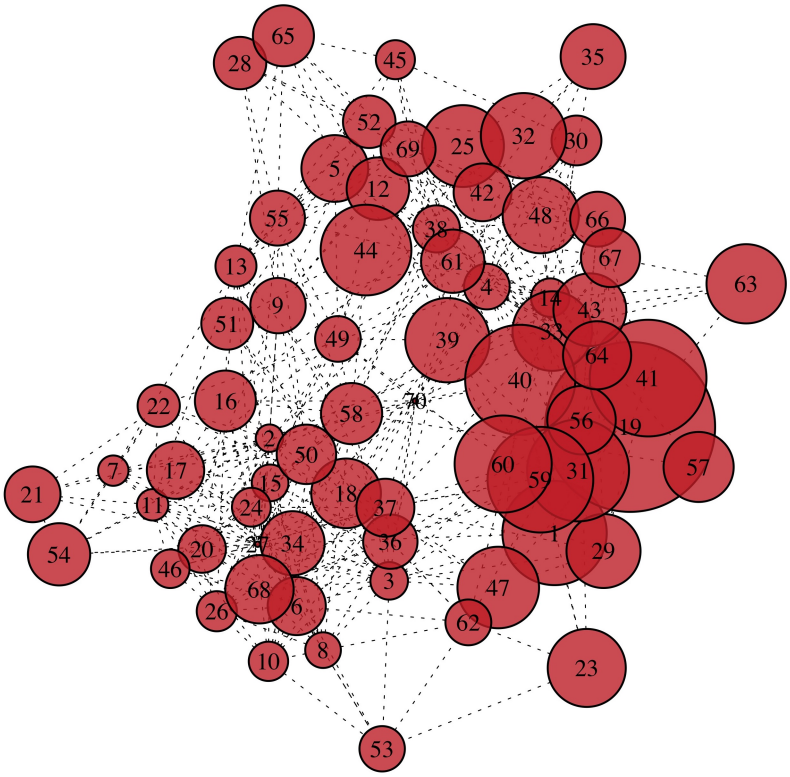
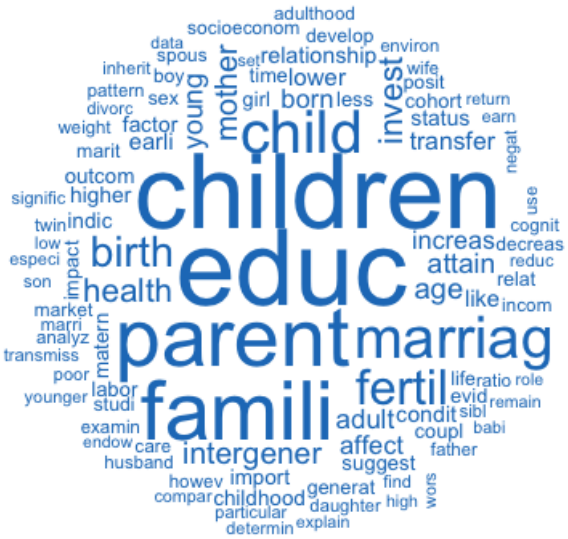


Figure A. 7 : Connectedness between topics and the female authors documents/abstracts in each topic. Extended sample with P&P articles.

it is easier to see that the latent topics go beyond standard research fields. In particular, Figure A.8 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 in related with family economics and topic 19 with gender discrimination.



(a) Topic 41

(b) Topic 19

Figure A. 8 : Topic Word Clouds in the extended sample with P&P articles