EC331 - Research In Applied Economics
Project Outline

Gender differences in sorting: Measurement and decomposition of productivity in academia

Milan Makany
u2048873

December 30, 2022

Project Outline

Research Question

Is there a gender differential in the sorting of academics?

Motivation and Objective

The primary motivation of this research is to analyse gender differences in academia by providing insight into differences in sorting – a discrepancy in high-productivity academics reaching high-productivity institutions. This is important for both societal and efficiency reasons, this research focusing on the latter. If the gap exists, it can potentially be explained by discrimination or an overall divergence in preferences.

In order to perform econometrics analysis I require a consistent cross-field measure of quality. This is vital as researchers have different publication patterns across fields and papers in journals of similar stature might have different impacts. I build on scientometric literature to construct a uniform measure of contribution using machine learning methods. To my knowledge no systematic analysis of this sort has been published.

Once I have a consistent measure of contribution, I deconstruct academic- and department-effects using an AKM model (Abowd et al., 1999). I observe differences in the distribution of individual and department effects and analyse the sorting effect by field, and region.

Contribution

This research has two contributions. First, I propose a method of comparing the quality of research across fields. This is a useful tool for grant and promotion committees along with academics looking to analyse science. Second, the difference in sorting may help explain a part of the observed gender differences and highlight an inefficiency in how we organise science.

Methodology

Measure of Contribution

I test multiple machine learning methods, this is a short summary of a selected few. For a more comprehensive review see Kuhn and Johnson (2013, Chapter 19).

I construct the metrics described in Waltman (2016) to obtain more than 100 measures of research quality and use machine learning to find the optimal weight assigned to each measure. For this to be feasible I construct a labeled database of academics and an outcome variable *top_institution*, equal to 1 if the academic is affiliated with an institution in the 95th percentile after the first ten years of their career and 0 otherwise.

I normalise all the measures to the range [0,1] using Equation 1 and compute the vector of weights $\vec{w}_f$ for each field by solving the optimisation problem in Equation 2 using gradient descent. The only hyper-parameter of question for gradient descent is the learning rate which can be adjusted via trial and error.

$$contrib_i = \frac{contrib_i - min(contrib)}{max(contrib) - min(contrib)} \tag{1}$$

$$\min_{w_m} \quad (top\_institution_i - \sum_m w_m score_{im})^2 \quad s.t. \quad \sum_m w_m = 1 \tag{2}$$

I also construct the vector of weights using a Neural Network that allows for more complex connection between the measures of contribution. Figure 1 demonstrates how the weights in a hypothetical network look. The connections between the nodes (weights) are more visible if that connection is more important in explaining the outcome variable. The number and size of hidden layers along with other hyper-parameters can be optimised for the best results.

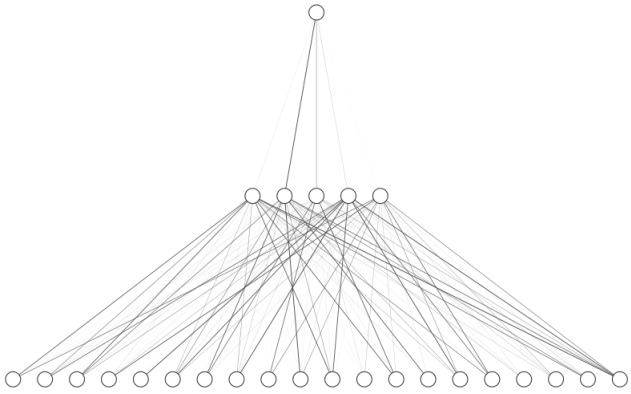

Figure 1: Example of a hypothetical neural network structure

I verify this approach with promotion data from Italy and Spain (Bagues et al., 2017) by computing the same weights with a new outcome variable *promoted*. If the obtained weights are significantly different, it is probable that the results are biased. In this case I will have to consider restricting the sample to certain fields which would reduce the external validity of the research.

Decomposition

I decompose the individual and department effects using an AKM model in Equation 3 (Abowd et al., 1999).

$$contribution_{ikt} = \alpha_i + \delta_t + \varphi_{kt} + \varepsilon_{ikt} \tag{3}$$

I estimate a simple two period model based on the specification proposed in Bonhomme et al. (2019), where α_i captures the individual and φ_k captures the department effect. I define event time as the number of years since first publication as a professor. This is difficult to pin down as in some fields it is quite common to publish during the course of a PhD and pursue post-doctoral positions.

In this two period model I first observe academics 5 years after their PhD working as Assistant Professors and then again 5 years later having worked as Associate Professors. (I have the data to estimate a more complex dynamic model, see Potential For Further Research)

Sorting defined as $cov(\alpha_i, \varphi_k)$ can be estimated separately for men and women. If the difference in sorting $cov(\alpha_i, \varphi_k|female = 1) - cov(\alpha_i, \varphi_k|female = 0)$ is negative then I show that high-productivity women are less likely to make it to high-productivity departments.

The sign of the φ_{kt} coefficients is also of interest. The interpretation of these coefficients is somewhat complicated. Department effects are a function of conditions provided by the department: facilities, support, networks, etc. I estimate these department effects at different intervals to analyse trade-offs associated with institutions providing different levels of support and assigning different amounts of administrative and teaching duty to academics in certain stages of their career.

Identification

The first condition for identification is to have enough variation in the data to estimate the fixed effects. I expect a large variance in publication patterns and quality, hence too little heterogeneity is not likely to be an issue.

The second requirement is a sufficient number of movers. It is reasonable to assume that the mobility in academia is sufficient for estimation, however this can be only verified by observing the data.

Third, institutions have to be connected to each other via movers (Abowd et al., 2002). If this fails using university-academic level data, I will group institutions with a k-means clustering algorithm (Bonhomme et al., 2019) or analyse a limited set of institutions and academics.

The fixed effects model also relies on the assumption of random timing in movements. This assumption is very strong and likely to be violated as moves in academia are usually associated with promotions and are endogenous. I will have to analyse the reasons for moving and ease this assumption.

Data Description

Literature Review

How do we conduct science? We must understand the organisation of discovery and the diffusion of knowledge. Stephan (2012) provides a detailed description of the scientific research process. They emphasize the variance in the process of production and collaboration across fields. The two key components of research are clearly defined: individual knowledge, persistence and ability (individual-effect captured by α_i) and environment, facilities, and networks (department-effect captured by φ_k). The AKM setup is well-suited for the analysis of production in science since the two main components of productivity are captured in the model. A possible extension of this research is to further decompose department effects and analyse what factors of production are most important in stages of an academic's career.

Waldinger (2012) shows that faculty quality has long-term effects on the productivity of PhD students and furthermore, high-quality scientists attract talent and funding (Waldinger, 2016; Azoulay et al., 2010). From an efficiency perspective this implies that it is beneficial to match high-productivity academics and that high-productivity departments provide the environment that allows for science to augment.

It is widely documented that there are persistent gender differences in science (Gasser and Shaffer, 2014; Larivière et al., 2013; Iaria et al., 2022). These inequalities transmit through multiple channels, one of which is the inefficiency they create. If high-productivity women are less likely to match to high-productivity environments and hence other high-productivity academics we do not utilise their potential and lose out on crucial peer effects. Ganguli (2015) shows that studies published before moving to a *better* department receive more attention after the move, implying that if a gap in sorting exists then women's publications are systematically under-utilised.

The importance of understanding gender dynamics in sorting is clear. In order to analyse it I use an AKM approach adopted by Card et al. (2016) to study the gender wage gap. The methodology they use to identify the gender differences can be perfectly applied for answering my research question. The novelty of my research is that the individual effects α_i can be interpreted in a more straight-forward manner since scientific productivity is public and directly observed. If a difference in sorting exists how do we interpret it? Is it a simple difference in preferences or are women systematically discriminated against? To answer this question is very ambitious and to do so I would need to infer preferences (for example like Sorkin, 2017) or construct some measure of discrimination.

To perform the decomposition I first have to create a metric that measures contribution consistently across fields. There is an extensive discussion in the Scientometrics literature about comparing the quality of publications by different authors in different fields. Radicchi et al. (2008) shows that citation patterns vary greatly across fields and claims that using a scaled measure of citations (Equation 4) produces a universally comparable metric across disciplines.

$$c_f = \frac{c_i}{\bar{c}} \tag{4}$$

Their claim is investigated by Waltman et al. (2012) using a larger sample of fields and publications. They only find circumstantial evidence to support the conclusion that this scaled metric allows for valid comparison across fields. Unfortunately, I do not find the results convincing and since I heavily rely on cross-field comparability, I use machine learning techniques outlined in Methodology.

Bibliography

Abowd, J. M., Creecy, R. H., & Kramarz, F. (2002, March). *Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data* (No. 2002-06). Center for Economic Studies, U.S. Census Bureau. Retrieved November 28, 2022, from https://ideas.repec.org/p/cen/tpaper/2002-06.html

Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, *67*(2), 251–333. https://doi.org/10.1111/1468-0262.00020
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00020

Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar Extinction. *The Quarterly Journal of Economics*, *125*(2), 549–589. https://doi.org/10.1162/qjec.2010.125.2.549

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, *107*(4), 1207–38.

Bonhomme, S., Lamadon, T., & Manresa, E. (2019). A Distributional Framework for Matched Employer Employee Data. *Econometrica*, *87*(3), 699–739. https://doi.org/10.3982/ECTA15722
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15722

Card, D., Cardoso, A. R., & Kline, P. (2016). Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women *. *The Quarterly Journal of Economics*, *131*(2), 633–686. https://doi.org/10.1093/qje/qjv038

Ganguli, I. (2015). Immigration and Ideas: What Did Russian Scientists "Bring" to the United States? *Journal of Labor Economics*, *33*(S1), S257–S288. https://doi.org/10.1086/679741

Gasser, C. E., & Shaffer, K. S. (2014). Career Development of Women in Academia: Traversing the Leaky Pipeline. https://doi.org/10.13016/M2Z892F9H
Accepted: 2017-09-20T12:32:57Z

Iaria, A., Schwarz, C., & Waldinger, F. (2022, June 30). Gender Gaps in Academia: Global Evidence Over the Twentieth Century. https://doi.org/10.2139/ssrn.4150221

Kuhn, M., & Johnson, K. (2013, May 17). *Applied Predictive Modeling* (1st ed. 2013, Corr. 2nd printing 2018 edition). Springer.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213. https://doi.org/10.1038/504211a

Moed, H. F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer Science & Business Media.

Priem, J., Piwowar, H., & Orr, R. (2022, June 16). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. https://doi.org/10.48550/arXiv.2205.01833

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272. https://doi.org/10.1073/pnas.0806977105

Sorkin, I. (2017). The Role of Firms in Gender Earnings Inequality: Evidence from the United States. *American Economic Review*, *107*(5), 384–387. https://doi.org/10.1257/aer.p20171015

Stephan, P. (2012). How Economics Shapes Science. In *E-BOOK GESAMTPAKET / COMPLETE PACKAGE 2012*. Harvard University Press.

Waldinger, F. (2012). Peer effects in science: Evidence from the dismissal of scientists in Nazi Germany. *The review of economic studies*, *79*(2), 838–861.

Waldinger, F. (2016). Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge. *Review of Economics and Statistics*, *98*(5), 811–831.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*(1), 72–77. https://doi.org/10.1002/asi.21671
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21671

Appendix

Potential For Further Research

Model Extensions

The network effects at institutions are of key interest. I estimate the model in Equation 5 where I control for the network of individuals. Close network defined as co-authors and broad network as colleagues. Note that close network has the subscript i because it depends on the specific person but the broad network depends on the department (excluding academic i) hence the k subscript.

$$contribution_{ikt} = \alpha_i + \delta_t + \varphi_{kt} + net\_close_{it} + net\_broad_{ikt} + \varepsilon_{ikt} \tag{5}$$

These network effects might be vital to explaining the difference in sorting. If women benefit disproportionately little from their broad network (colleagues) they have less incentives to sort into high-productivity departments and rather focus their efforts on building their close network.

This exercise is quite feasible as Bonhomme et al. (2019) provide the methodology for estimating more complex dynamic models and could spur more research questions.

Evolution of Fields

Having found the best method for constructing a field-consistent measure of contribution I can use the algorithm to obtain the appropriate weights for fields throughout time (recalculating the weights each decade or so). I expect that more recently established (young) fields such as economics place a higher weight on the biggest contribution (best paper published by an author), while older fields might place higher weights on the number of publications.

Tracking the evolution of fields this way might provide insight for new academics on what to focus on at the start of their career if they want to be successful in the future.

Inequality of Opportunity

In a similar manner as in Evolution of Fields I can also track the probability of making it to a top institution, given the academic is a top-productivity individual (based on the decomposition).

$$P(Top\_Inst|Top\_Academic, Female = 1) - P(Top\_Inst|Top\_Academic, Female = 0) \tag{6}$$

I define the gender gap in opportunity in Equation 6. This can be estimated for multiple time periods, where time can be defined in absolute terms or the maturity of a field (time since first journals established). The same analysis can be done using the AKM model in Equation 3 to track the dynamics of sorting. This descriptive analysis could be useful in understanding how gender dynamics evolve in academia across and within fields.

Understanding Research

This study's primary goal is to provide insight into the *machinery* of academia. However, in the process of constructing measures of contribution I approach the issue from a purely statistical perspective. I find it important to understand why certain metrics carry more weight than others;

is this an accurate reflection of how science contributes to society or rather a reflection of how academics judge each other in the process of promotions. There is a short review of the difference between bibliometrics and scientometrics in the Literature Review and the Appendix.

Rankings

Comparing the department fixed effects to the already existing rankings is an interesting exercise. This tool could be useful for academics when facing a decision of where to move. Based on the predictions of this model they would be best off going to a place with a very high estimated department fixed effect φ_k.

Data Description

I use publication data from OpenAlex, "an index of hundreds of millions of interconnected entities across the global research system" (Priem et al., 2022). They provide an open-access database of publications, author affiliations, journals, the number of citations since publication (in some cases even the citations of the working paper, before publication) and the works cited by the authors.

The official documentation of the objects in OpenAlex

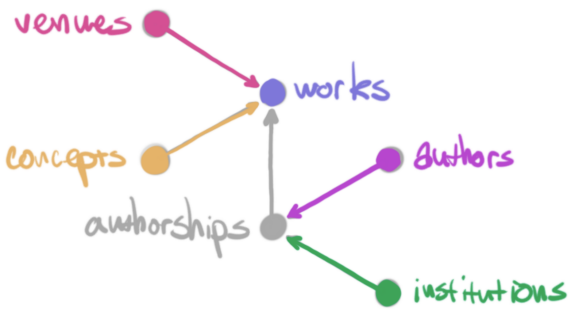

Figure 2: Connection of OpenAlex Objects

OpenAlex has data of 6 different types shown in Figure 2. The primary object of interest for my research is Works. This object contains data from all other objects and has references to the other objects in case I need some additional data not contained in Works. The data is stored in JSON format (equivalent of dictionaries in Python), a list of key-value pairs. Samples of objects can be found in Table 1.

Authors	https://api.openalex.org/people/A2056708386
Works	https://api.openalex.org/works/W3121763541
Venues	https://api.openalex.org/journals/V203860005
Institutions	https://api.openalex.org/institutions/I39555362

Table 1: OpenAlex Objects

Data Collection and Extraction

There are multiple methods for extracting data from OpenAlex. They provide a public API with a rate limit of 10 requests per second and they provide snapshot saves of the entire compressed database from AWS.

I copied a snapshot of the database (323 GB) onto an external SSD and then to the University's Scientific Computing Research Technology Platform (SCRTP) cluster computer. From this compressed database I extract the data using Python and then process it using either Python or Stata.

If the snapshot turns out to be incomplete or unable to process the API is the second best option. Using Python I can send a series of HTTP requests to download publication data. This process is more time consuming, however the time allocated to programming would be reduced.

Bibliometrics v Scientometrics

Bibliometrics in general is more about understanding what quality is rather than simply looking for ways to measure it. Scientometrics literature focuses more on creating metrics which measure the academic's contribution based on observable statistics.

Example of a bibliometric approach in Moed (2006, p. 151)

> *First, one should collect documents containing statements of scholars in the field under study on how assessment of research performance should be conducted, and, of course, on how it should not be conducted. Earlier reports of peer review committees evaluating scholars in the field constitute a fruitful basis for such an inventory. The bibliometric investigator should identify the main aspects of research quality involved, issues that were raised, problems that remained unsolved, operationalisations that were applied or rejected. Secondly, scholars from the field should be involved in all stages of the study. They should be stimulated to propose or develop – even preliminary – classification systems, and to structure their own research output accordingly.*

Example of a scientometric approach in Waltman (2016, p. 20)

> *In the calculation of citation impact indicators, citations are sometimes taken into account only within a specific time period after the appearance of a publication, the so-called citation window. Adopting a certain citation window may cause both publications and citations to be excluded from the calculation of citation impact indicators. For instance, suppose we require publications to have a citation window of at least five years. For recent publications it is not possible to have a five-year citation window, and therefore this requirement implies that recent publications cannot be included in the calculation of citation impact indicators.*

In general I find the bibliometric approach more appealing as it aims to provide a deeper understanding of research and how to define contribution, however for the goals of this project the scientometric approach is more important. A natural extension of this study is to understand why certain scientometric indicators are more important than others and this is where methods from the bibliometric approach should be adopted.