Gender issues in fundamental physics: a bibliometric analysis

Alessandro Strumia

Abstract

We analyse bibliometric data about fundamental physics world-wide from 1970 to now extracting quantitative data about gender issues. We do not find significant gender differences in hiring rates, hiring timing, career gaps and slowdowns, abandonment rates, citation and self-citation patterns. Furthermore, various bibliometric indicators (number of fractionally-counted papers, citations, etc) exhibit a productivity gap at hiring moments, at career level, and without integrating over careers. The gap persists after accounting for confounding factors and manifests as an increasing fraction of male authors going from average to top authors in terms of bibliometric indices, with a quantitative shape that can be fitted by higher male variability.

1 Introduction

This paper originates from an observational opportunity: for the first time sociological issues in fundamental physics can be studied using the public InSpire database, that accumulated bibliometric data about fundamental physics world-wide from $\sim$ 1970 to now [InSpire(2010)]. Fundamental physics is a sub-sector of physics that deals with the fundamental aspects of the field and that presently focuses mostly on particle physics, cosmology and astrophysics, from an experimental and theoretical point of view.

Such bibliometric data are being used to study various aspects of the field. Like other Science, Technology, Engineering, and Mathematics (STEM) fields, physics exhibits persisting gender differences that we try to characterise and understand in the present paper. The bibliometric approach relies on large amounts of objective quantitative data about papers, authors, citations and hires. Having a large amount of new data we will follow a data-driven approach. Enough statistics is sometimes needed to reveal effects, and to go beyond simple counting by devising dedicated analyses that target specific questions. We can do this, as we have the full database, not just access to some pre-defined metrics.

While a vast literature studied gender differences in STEM, no previous studies specifically

focused on fundamental physics: the present study will fill this gap.[1] A main theme is understanding why women remain under-represented in STEM fields, a worldwide phenomenon persisting since decades, despite interventions on its alleged social causes [Stoet et al.(2018)]. A limitation of bibliometric analyses is that authors start being scientifically active roughly at PhD level: in physics (as in other STEM fields) a low female representation is already present at this entry level of bibliometric data. Earlier phases need to be explored with other tools.

An important clue is that a similar gender difference already appears in surveys of occupational plans and first choices of high-school students [Xie et al.(2003), Ceci et al.(2014)]. This is possibly mainly due to gender differences in interests [Ceci et al.(2014), Su et al.(2009), Lippa (2010), Hyde (2014), Su et al.(2015), Thelwall (2018)b, Stoet et al.(2018)]. Gender differences in relative attitudes (girls with high math ability tend to also have high verbal ability) also contribute to student choices [Wang et al.(2013), Ceci et al.(2014), Stoet et al.(2018)]: most of the gender gap in student intentions to study math disappears after taking into account their math versus reading difference in PISA scores, while absolute results (boys outperform girls in math, girls outperform boys in reading) are much less able to explain the gender gap [Breda et al.(2019)].

Coming to the later phase that can be studied by experiments and by bibliometrics, initial small-scale experiments and anecdotal reports suggested biases against hypothetical female applicants (see e.g. [Wenneras et al.(1997), Moss-Racusin et al.(2012)]; see also [Eaton et al.(2019)]). These findings have not been supported by more recent larger-scale experiments (see the review in [Ceci et al.(2011)] and [Williams et al.(2017), Ceci et al.(2015)]). [Milkman et al.(2015)] sent letters by fictional students seeking research opportunities to professors and measured their response rate. The result of this social experiment performed in the US was that, looking at gender in isolation (rather than at 'women and minorities'), female students received slightly more responses in public schools (the majority of the sample) with respect to men in the same racial group.

Experiments about hypothetical applicants often miss key elements of real applications, mostly based on scientific results reported in publications and evaluated by scientists who work in the same sub-field. No significant biases have been found in examined real grant evaluations [Ceci et al.(2014), Marsh et al.(2011), Ley et al.(2008), Mutz et al.(2012)] and referee reports of journals [Borsuk et al.(2009), Ceci et al.(2014), Edwards et al.(2018)]; the gender composition of applicants [Way et al.(2016)] and panels [Abramo et al.(2018)] has little effects. Real hires show a higher success rates among women [National Research Council, Wolfinger et al.(2008), Glass et al.(2010), Ceci et al.(2014)], especially in those STEM fields where women

[1] Various studies focused on discriminations as a possible source of gender differences. Small samples of female physics students were interviewed by [Barthelemy et al.(2016)] and [Aycock et al.(2019)]. The [NASEM report (2018)] focused on the University of Texas System, finding that 17% (13%) of female (male) students in Science reported "sexist hostility". However, the [NASEM report (2018)] also found that a higher 45% rate of "sexist hostility" is reported by female students in Medicine, a field with a negligible gender gap in participation. Only a few % of male and female STEM students reported more serious problems (fig. 3.3 of [NASEM report (2018)]). As "evidence of direct discrimination is limited", alternative interpretations of the gender representation difference in STEM have been considered and "many scholars now emphasize the role of gender differences in preferences, self-concept and attitudes" [Breda et al.(2019)].

are less represented [Ceci et al.(2014)]. Bibliometric attempts of recognising higher merit [Ceci et al.(2014)] found that male faculty members write more papers [Xie et al.(1998), Levin et al.(1998), Fox (2005), Abramo et al.(2009), Lariviere et al.(2013), Way et al.(2016)] (see also [Holman et al.(2018), Thelwall (2018)]), predominate among first and last authors (prestigious in some fields) and in single-authored papers [West et al.(2013), Jagsi et al.(2006)]. Such gender productivity gap persists after accounting for confounding factors such as seniority [Ceci et al.(2014), Caplar et al.(2017), Moldwin et al.(2018)]. Consistently with these results, [Witteman et al.(2019)] found that female grant applications in Canada are less successful when evaluations involve career-level elements. Some studies observed a small group of extremely productive, mostly male, 'star authors' [Bordons et al.(2003), Abramo et al.(2009)b, Abramo et al.(2015)]. A smaller 'leaky pipeline' rate of female authors is observed in STEM fields than in other fields with higher female representation [Ceci et al.(2014)]. Looking at PLOS medical journals, where each author declares his/her role (analysis, design, material, perform, write), [Macaluso et al.(2016)] found that women were more involved in performing experiments, and men more involved in the other roles.

This paper is structured as follows.

In section 2 we describe how we identify the gender of authors; how we obtain lists of hires; how we combine citations to define bibliometric indicators which can be used as reliable proxies for scientific merit, being significantly correlated to human evaluations such as scientific prizes; how we deal with the confounding factor due to the gender evolution of the field.

In section 3 we present findings that exhibit interesting gender differences. New authors today appear with roughly 4:1 male:female proportion, with order one variations in different countries. We find that this entry difference in representation is negligibly affected by hiring, consistently with [Ceci et al.(2014)]. As we use citations, in section 3.1 we first verify that male (M) and female (F) authors cite in the same way. We achieve this by defining a gender asymmetry in citations sensitive only to a differential gender bias, not to gender differences in number or productivity of authors. n section 3.2 we then compare reliable bibliometric indices based on citations finding that F authors are hired with indices which are, on average, not higher than those of M authors. Section 3.3 finds a productivity gap consistent with previous studies. This new difference is quantitatively studied in section 3.4 finding that the M fraction progressively grows going from average to top-authors, consistently with [Bordons et al.(2003), Abramo et al.(2009)b, Abramo et al.(2015)]. In section 3.5 we consider self-references, finding no new gender differences. Data are made available in Appendix A. As bibliometric data are influenced by a complicated background of social and historical accidents in the supplementary section S1 we show that the results above persist after taking confounding variables into account. Statistical details are presented in the supplementary section S2. Interpretations of the data are discussed in the Conclusions, section 4.

2 Methods

The public InSpire database [InSpire(2010)] maintained by CERN and other institutions offers a picture of fundamental physics world-wide from $\sim$ 1970 to now. InSpire gives data on about 1.3 million of scientific papers, 30 million of references, 71104 identified authors in 7 thousands of institutes and 6 thousands of collaborations. InSpire individually identified all authors (except occasional authors), solving the problem of name disambiguation. We expect that the database is negligibly affected by direct or indirect gender bias. Indeed, the database provides essentially full coverage of the scientific literature within "fundamental physics". Since decades this is a self-contained highly specialised subject [Sinatra et al.(2005)], so that "boundaries" of the data-base play a minor role. Papers in InSpire include all those published in some categories of the pre-print bulletin arXiv and in some journals with topics considered relevant for fundamental physics. Human intervention is minor, such as adding extra papers considered relevant. Only a minor fraction of authors occasionally work in other fields or arrive from other fields. A possible gender difference in multi-disciplinary attitudes is thereby expected to negligibly impact our subsequent discussion. On the other hand, various authors work on multiple topics within the field, that cannot be sharply sub-divided.

InSpire does not provide gender information: in section 2.1 we describe our procedure to infer gender from full names and nationality of the authors. In section 2.2 we describe how we obtain lists of hires in fundamental physics world-wide. Section 2.3 motivates the bibliometric index that we will use to indicate scientific merit.

2.1 Name-gender association

We need to infer gender from names in an accurate and complete way.[2] Three main problems are encountered. First, the InSpire database provides only name initials for about 13% of the authors. These are mostly authors with little impact, as defined by any index. Second, some names like Nicola are "ambiguous": they correspond to different genders in different countries. Third, some authors have unusual names. The `Mathematica` machine learning function `Classify` [Wolfram et al.] uses information about the first name only and leaves about 40% of authors with unclassified gender.

We tested two approaches, in order to determine their strengths and to choose the best combination:

1. First, we run the on-line Ethnea [Torvik et al.(2016)] tool, which uses the full name (first and family name) to infer both gender and ethnicity. Ethnea leaves 26% of the authors with unclassified gender.

2. Second, for each author we extract a "guessed" nationality from the earlier affiliations in his/her papers and use it to disambiguate "ambiguous" names. The obtained list of first names and nationalities is matched to a database of names and countries from the

[2]Some U.S. astronomers "discourage" adopting this "quantitative methodology", seen as "epistemically violent" and "discriminatory" [Rasmussen et al.(2019)].

Worldwide Gender-Name Dictionary (WGND) [WGND]. This database contains 175917 names with their associated countries. About 70% of authors have "unambiguous" names that are present in the WGND. Authors with "ambiguous" names present in the WGND are matched using the nationality inferred from their earliest affiliations. The size of this subset of authors is $\sim 3\%$ of the total, and the uncertainty induced by this procedure is below the percent level. About 0.1% of the authors have "ambiguous" names and no nationality information: we match them to the most common gender corresponding to their name, defined as the one used in the largest number of countries. 23% of the authors remain unclassified.

The results discussed in the following are affected in a minor way using the Ethnea or the WGND classification. By comparing them we see that that the Ethnea classification is less complete, leaving unclassified more authors with unusual names. On the other hand, the WGND classifications leads to some authors with misidentified gender, typically arising due to a misidentification of their nationality. We find different genders for 1.8% of all identified authors; the percentage grows up to 5% among Chinese,[3] Indian and Korean authors, and decreases down to 1% among European authors.

As a best choice, we adopt the Ethnea classification whenever available, and the WGND classification otherwise. Furthermore, we selected a thousand of top-cited authors in different time periods and systematically verified and correctly assigned their gender with no errors, using information available on internet.

2.2 Hiring

InSpire is integrated with HepNames, a data base with biographical information about the various authors, including papers, affiliation history, experiments they participated in, PhD advisor, graduate students. As an example the internet page inspirehep.net/author/profile/A.Strumia.1 shows the profile of the present author. A user interface allows researchers to create and update HepNames records for themselves and for other authors, providing precise career information on a voluntary basis (to be validated by the InSpire team). Furthermore, large collaborations systematically provide complete author information upon submission of documents through a dedicated format.

From HepNames we obtain a data-base of about 10000 first hires in fundamental physics world-wide, including dates and disambiguated institutions. Such InSpire hires might be biased if those F and M authors who need to self-report their data tend to do this differently. While InSpire is a widely used tool in the community, integrated with a job announcement system, funded by multiple official institutions and endorsed in Reviews of Particle Physics, occasional authors might not use it.

We therefore complement InSpire hires by computing unbiased "pseudo-hires" defined as follows. For each paper, we have a list of disambiguated affiliations of each author. We consider

[3]Gender can be reliably extracted from Chinese names only when they are written in Chinese characters: this information is not always provided by InSpire.

an author as $p$yr-hired when he/she starts writing papers with the same affiliation for at least p years. Using this definition we obtain a database of about 40000/19000 5/10yr-first hires from 1960 to 2013/2008 (64000/23000 including multiple hires for the same author). However, in this way we cannot obtain a precise hiring date for the sub-set of authors hired by the same institution to which they were previously affiliated.

Thereby we will use InSpire hires when a precise hiring date is more important than increased statistics, and pseudo-hires in the opposite situation, when a full coverage is more important than a precise timing. In any case, the other sample will be used as a control sample.

2.3 Bibliometrics

We here motivate the use of appropriate bibliometric indicators as a proxy for what is commonly considered as scientific merit.

Various authors studied what citation counts do measure.

At theoretical level, two main models have been proposed. According to the normative interpretation, scientists primarily cite to give credit. Then, a bibliometric index provides a valid proxy of scientific merit, especially when highly correlated with scientific prizes or other human evaluations of scientific merit. According to the social-constructivist interpretation citations are instead primarily a social persuasion tool; the concept of scientific merit itself is questioned: as reviewed in [Bornmann et al.(2008)] "scientific knowledge is socially constructed through the manipulation of political and financial resources and the use of rhetorical devices". According to this point of view, citation counts could be correlated to prizes simply because both reflect social status. Some prizes in physics require established scientific results, others are awarded following rules that leave more space to sociological distortions.

Observational works supported the normative interpretation at high aggregation level [Bornmann et al.(2008),Tahamtan et al.(2019)]; personal factors important at individual level average out when considering many authors. Our data provides extra evidence in this direction: for example authors of top-cited papers tend to be younger, rather than senior powerful scientists.

Citation counts are surely influenced by some confounding factors [Bornmann et al.(2008)]: the citation intensity depends on time and field (our indicator will compensate for this), on language (essentially all physics literature is in English), on accessibility (physics literature is freely available on the pre-print bulletin arXiv since 1995), on collaboration size.

Collaboration size is a big issue in fundamental physics, due to the presence of very large (up to 3000 authors) and very productive (up to 6000 papers) collaborations, mostly in high-energy experimental physics. Because of this main reason traditional metrics (such as citation counts, h index, paper counts) now fail to provide reasonable proxies for scientific merit in fundamental physics [Strumia et al.(2018)]. Signing more papers that what one can read stretches the concept of authorship [Birnholtz et al.(2006)]. At quantitative level, the problem is that the contribution of one big collaboration overwhelms the data-base (1.3 million of papers), if 6000 papers are counted as $3000 \times 6000 = 1.8$ million.

This situation can be corrected by "fractional counting" [Hooydonk (1997),Perianes-Rodiguez et al.(2016),Leydesdorff et al.(2016)]: a fraction $1/N_{\rm aut}$ of each paper (rather than the full paper)

is equally attributed to its $N_{\rm aut}$ authors, as appropriate for an intensive quantity. All authors, including first and last authors, are treated on equal footing because authors are usually sorted alphabetically in fundamental physics, unlike what happens in other fields.[4] Thereby there is no way of telling who contributed what to multi-authored papers. When huge collaborations are involved there is no warranty that each author contributed to each paper. Despite this, data show that the total fractionally-counted bibliometric output of collaborations scales, on average, as their number of authors [Rossi et al.(2019)], suggesting that large collaborations form when is scientifically needed and that gift authorship does not play a large role.

Fractional counting of citations already provides one simple acceptable indicator. We improve on it by using the closely related number of "individual citations"

$$N_{\rm icit} = N_{\rm cit}/N_{\rm aut}N_{\rm ref} \tag{1}$$

(summed over all citing papers, as precisely defined in [Strumia et al.(2018)]) that gives reduced weight to citations coming from papers with a larger number $N_{\rm ref}$ of references. This refinement addresses the issue of normalization between different fields and times [Zitt et al.(2008)] (see [Waltman (2015)] for a review and extra references): papers in sectors with a higher rate of publications (such as phenomenology in fundamental physics) tend to receive more citations; for the same reason these papers also tend to have more references. Thereby, dividing by the number of references tends to give a common normalization to different fields, without needing a field classification system. Indeed, the average number of individual citations received by papers in any field disconnected from other fields is 1. As a test that this concept works in practice in the InSpire database, we computed the average number of citations, of references and of individual citations of papers within the main theoretical fields defined by arXiv (hep-th, hep-ph, gr-qc, nucl-th, astro-ph, hep-lat[5]), finding that the dispersion in $N_{\rm icit}$ among different fields is reduced down to 8%, more than twice smaller than the dispersion in $N_{\rm cit}$ or $N_{\rm ref}$. Similar results are found considering different times: the field grew with time, such that newer papers receive more citations and have more references, roughly in proportional amounts.

Individual citations have the following meaning: an author who wrote $N_{\rm icit}$ fractionally-counted papers of average impact in his/her field received $N_{\rm icit}$ individual citations. Table 7 of [Strumia et al.(2018)] lists the 50 physicists who received most individual citations together with their scientific prizes. Physicists can read their names and consider if $N_{\rm icit}$ is dominantly influenced by scientific achievements or by social constructivism. For practical purposes, an indicator provides an acceptable proxy of scientific merit if scientific merit positively affects the index more than confounding variables. A full or large correlation with scientific merit improves the sensitivity of the analysis, but some effects can be large enough that a fine sensitivity is not needed to reveal them.

The use of bibliometric indices based on citation counts as a proxy for scientific merit comes

[4]The first author is not alphabetically sorted in 6% of the multi-authored papers in the hep-th arXiv bulletin, 13% in hep-ph and hep-ex, 18% in hep-lat, 25% in gr-qc, 44% in astro-ph. In more papers the author highlighted as first might accidentally be also alphabetically first.

[5]Experimental papers form a separate category, as they tend to have many co-authors and to receive more citations.

with limitations and dangers. Some citations are given for negative reasons. On short timescales citations are more influenced by visibility, and some authors engage in boosting their citation counts in various ways: large collaborations, many references, self-references, citation networks, salami slicing into minimum publishable units... Individual citations are not boosted by the first two strategies. As we are concerned with gender differences, it is reassuring that section 3.5 will find no extra significant gender differences in self-referencing.

Since "when a measure becomes a target, it ceases to be a good measure", in the supplementary section S1 we consider a metric based on citations more different from common targets, which is not enhanced by the latter three strategies. The 'CitationCoin' ₡ is defined as the difference between the number of received and given individual citations (up to a correction factor that prevents systematically negative contributions from recent papers), such that it is not affected by self-citations not by networks of circular citations [Strumia et al.(2018)]. Authors that write too many poorly cited papers can even have a negative ₡ score.

Bibliometric indicators measure the average opinion of the community: while all opinions can be wrong, a better possibility could be relying on the opinion of top-authors. This is done by metrics based on the PageRank algorithm (such as those discussed in [Pinski et al.(1976), Chen et al.(2007), Ma et al.(2008), Radicchi et al.(2009), West et al.(2014), Strumia et al.(2018)]). This is studied in the supplementary section S1, where for completeness we also consider the widely used but naive bibliometric indicators based on paper counting and on the average number of citations per paper.

In practice, the differences in bibliometric indices among authors are so large that log-scale plots will be appropriate and refined metrics only make minor differences. We use individual citations $N_{\rm icit}$ because this metric is simpler and closer to the commonly used number of citations $N_{\rm cit}$, while allowing to meaningfully deal with experimentalists, theorists and astrophysicists, by compensating for the vastly different typical number of co-authors $N_{\rm aut}$ of papers produced by these communities.

2.4 The age confounder

The fraction of female authors in fundamental physics significantly increased with time, producing demographic gender differences (female authors are on average younger than male author) that act as a confounding factor to our later analyses. Apparent gender differences can just be age differences. Career-integrated indices tend to favour senior authors, and indices based on single papers tend to favour younger authors. Since age is a significant confounder, we will compensate for the different time evolution $N^{\rm start}_{F,M}(t)$ (number of F and M authors that produced their first paper during year t) by assigning to each author A a weight proportional to

$$\frac{N^{\rm start}_F(t_A) + N^{\rm start}_M(t_A)}{2N^{\rm start}_G(t_A)} \tag{2}$$

where t_A is the date of his/her first paper and G is his/her gender. This is equivalent to selecting every year a random sub-set of new authors (respecting the time evolution of the total number

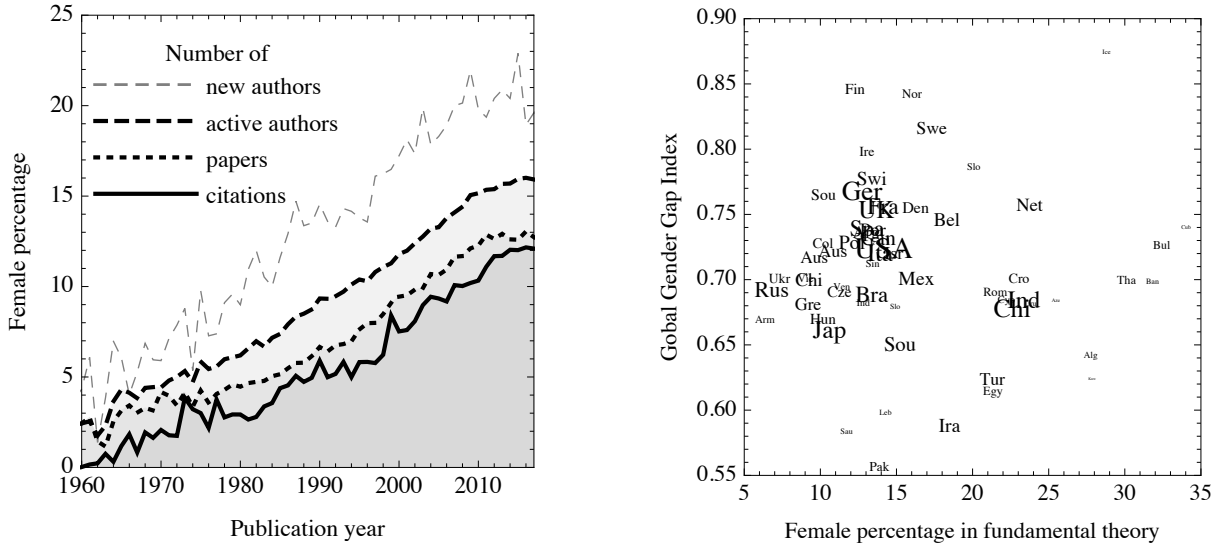

Figure 1: **Left**: *percentage female contribution to the number of new authors, of authors that wrote at least a paper during the year, to the number of fractionally-counted papers, and to the number of received individual citations, considering the papers written each year. Citations are counted based on the year of the cited publication.* **Right**: *the percentage of F authors in fundamental theory is not positively correlated with the Global Gender Gap Index of the country [World Economic Forum].*

of authors) such that M and F authors are numerically equal, and averaging over the possible choices.

3 Results

Among the 71104 authors in fundamental physics listed in the InSpire data-base we identified 49860 male and 9205 female authors. 16% of authors with identified gender are classified as female, and wrote 10% of the fractionally-counted papers receiving 7% of the individual citations. These raw numbers, meant only to give a first rough idea of the field, are affected by a variety of historical accidents.

As documented in [Strumia et al.(2018)] the field significantly expanded: due to increased publication intensity about half of citations have been given after 2000, so that metrics based on citations favour recent authors (our metric $N_{\rm icit}$ automatically compensates for publication intensity). Furthermore, the F percentage grew with time as shown by the raw data in the left panel of fig. 1.

The right panel shows that, within the countries that most contributed to fundamental physics, the female fractions range between 7% and 23%. It is interesting to explore if the female fraction is correlated with the Global Gender Gap Index (GGGI) of the countries [World

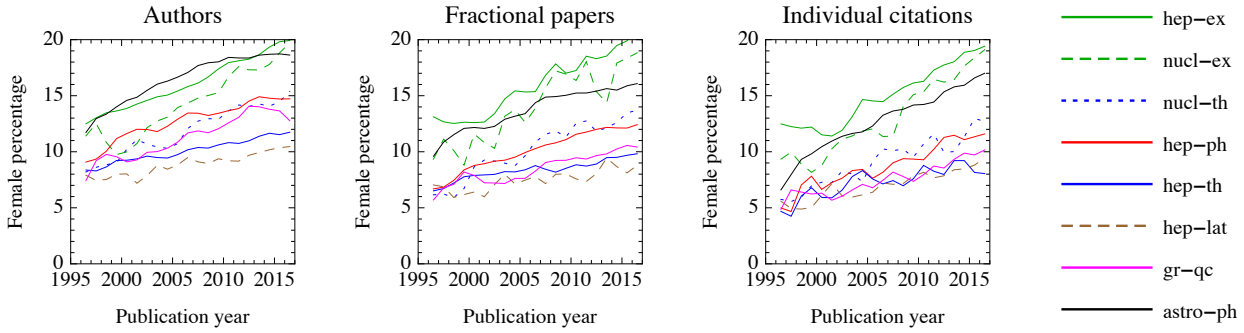

Figure 2: *As in fig. 1, showing after 1995 the result within the main arXiv categories, plotted as colored curves: experimental categories include hep-ex (high-energy experiments) and nucl-ex (nuclear experiments). Theoretical categories include hep-ph (high-energy phenomenology), hep-th (high-energy theory), hep-lat (lattice), nucl-th (nuclear theory); gr-qc (general relativity and quantum cosmology) is mostly theoretical, although it includes some experiments. Finally astro-ph contains astrophysics and cosmology.*

Economic Forum] which measures the gap between women and men in education, politics, health, economy, as this is a possible cause of the low female representation. The GGGI ranges between 0 and 1, with 1 indicating parity or a gap in favour of women (as the GGGI ignores imbalances to the advantage of women). The right panel of fig. 1 shows that the female fraction is not positively correlated with the GGGI, as similarly observed among students in STEM [Stoet et al.(2018)].

Fig. 2 shows that the female percentage is a factor of 2 higher in sub-fields dominated by large experimental collaborations than in theoretical fields.

Clearly, the field and its gender composition evolved in the past 50 years. While describing such changes from a bibliometric point of view is an interesting subject, we try focusing on general features which emerge from the complicated background of social factors. This will need taking into account possible confounding variables, by studying sub-periods and sup-topics or by trying to compensate for the above variations.

3.1 Citations

We want to investigate if citations are influenced by the gender of the cited authors, searching for a possible different tendency of the two genders to cite more often a given gender.

In principle, complete information could be extracted by comparing 'how citations are' with 'how citations would be' in the absence of gender discrimination. In practice this strategy needs a theoretical model of citations, but such models are affected by questionable systematic issues. One can try controlling for main factors (such as different numbers of M and F authors, different average seniorities, regional differences, etc), but reality can contain more complicated effects

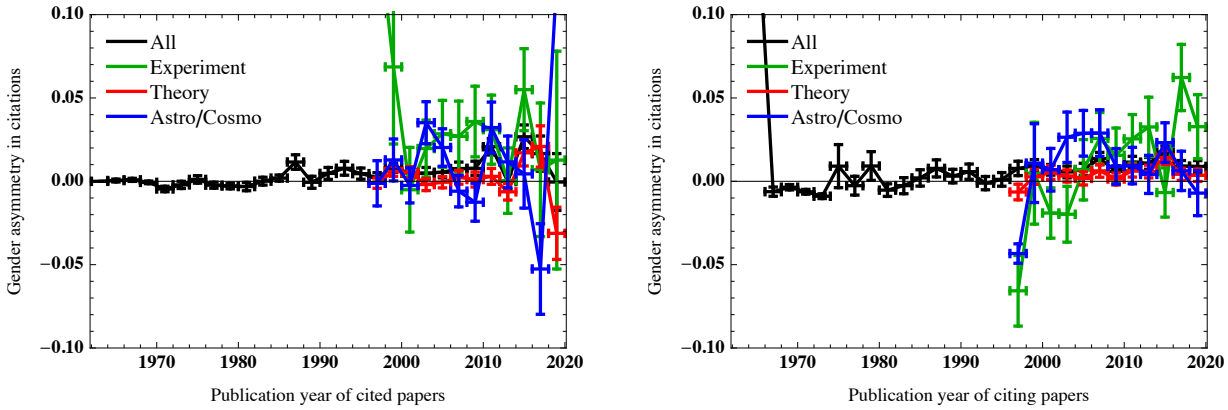

Figure 3: *Time evolution of the gender asymmetry defined in eq. (3); $A > 0$ ($A < 0$) signals same-gender (opposite-gender) preference.* **Left***: As function of the publication year of the cited single-author papers.* **Right***: As function of the publication year of the citing single-author papers. After 1995 we also show the asymmetry in different sectors of fundamental physics, based on their arXiv categories: theory (hep-ph, hep-th, hep-lat, nucl-th, and gr-qc), experiment (hep-ex and nucl-ex) and astrophysics (astro-ph). The bin 2018-20 only uses data available up to mid 2018.*

such as different scientific qualities. For example, [Caplar et al.(2017)] claim (consistently with our later findings) that papers in astronomy written by F authors are less cited than papers written by M authors, even after trying to correct for some social factors. After considering attributing the remaining difference to gender bias, [Caplar et al.(2017)] conclude "of course we cannot claim that we have actually measured gender bias".

We will follow a different strategy, which is often more useful in the presence of backgrounds that cannot be reliably modelled: we construct an asymmetry such that it is not affected by the backgrounds. The extracted information encoded in the asymmetry is reliable but partial, as we give up on the attempt of modelling the full citation process.

To start, we restrict our inquiry to the sub-sample of single-author papers with identified gender G, as these would likely be more strongly affected by a possible gender bias. We count $N^{\rm cit}_{G\to G'}$, the number of single-author papers with gender G citing single-author papers with gender G'. We compute the proportions $f_{G\to G'} = N^{\rm cit}_{G\to G'}/N^{\rm cit}_{G\to}$ dividing by the total numbers $N^{\rm cit}_{G\to} = \sum_{G'} N^{\rm cit}_{G\to G'}$, so $0 \le f_{G\to G'} \le 1$. From this we define the gender asymmetry as

$$A = f_{M\to M} - f_{F\to M} = f_{F\to F} - f_{M\to F} = \frac{1}{N^{\rm cit}_{M\to} N^{\rm cit}_{F\to}} \det \begin{pmatrix} N^{\rm cit}_{M\to M} & N^{\rm cit}_{M\to F} \\ N^{\rm cit}_{F\to M} & N^{\rm cit}_{F\to F} \end{pmatrix} . \qquad (3)$$

The first formula means that A is the proportion at which solo males cite solo male research more than solo females cite solo male research. The second formula means that A also is the proportion at which solo females cite solo female research more than solo males cite solo female research. So the gender asymmetry ranges between $-1 \le A \le 1$. The final formula shows

category	hep-ex	hep-ph	hep-th	hep-lat	nucl-ex	nucl-th	gr-qc	astro-ph
counts	2755	14627	15370	1762	1673	1258	6706	6733
A in %	-1.0 ± 1.7	0.5 ± 0.6	0.0 ± 0.7	-0.3 ± 2.2	6.0 ± 2.4	0.7 ± 2.2	0.5 ± 1.2	0.5 ± 1.1

Table 1: *Gender asymmetry A defined in eq. (3) computed restricting to single-author papers after 2010, in the arXiv categories defined in the caption of figure 3. The counts are the number of single-author papers in a given arXiv category cited by any single-author papers, not necessarily in the same category.*

category	hep-ex	hep-ph	hep-th	hep-lat	nucl-ex	nucl-th	gr-qc	astro-ph
counts/1000	115	421	270	42	22	44	90	285
A in %	0.0	0.3	0.5	1.0	1.0	0.5	-0.1	0.4

Table 2: *As in table 1, considering all papers after 2010.*

that A is symmetric under $M \leftrightarrow F$ permutations, with a property that makes it useful: A vanishes whenever citations are given without considering gender. $A > 0$ ($A < 0$) signals same-gender (opposite-gender) preference, althought more complicated patterns are possible: only one gender might have a particular preference for citing a given gender, or both might have a preference for opposite genders, or both might have a preference for the same gender, in different amounts. On the other hand, A is insensitive to a difference in the total number and in the average scientific quality of M and F authors (as quantified by the chosen indicator), as well as to a possible collective equal bias of both genders towards one gender, which corresponds to multiplying one column of the matrix above by a fixed constant.[6]

To better understand what the asymmetry measures, it is useful to compute its predicted value in a toy model of citations where $N_G^{\rm aut}$ authors of gender G cite with gender-dependent rates $p_{G\to G'}$ (for simplicity we ignore that some authors are more active than others, so that an effective number would be directly relevant). In this model $N^{\rm cit}_{G\to G'} \propto N_G^{\rm aut} N_{G'}^{\rm aut} p_{G\to G'}$ and the asymmetry equals

$$A \simeq \frac{N_M^{\rm aut} N_F^{\rm aut}}{(N_M^{\rm aut}+N_F^{\rm aut})^2} \det \begin{pmatrix} p_{M\to M} & p_{M\to F} \\ p_{F\to M} & p_{F\to F} \end{pmatrix} \tag{4}$$

in the limit where all $p_{G\to G'}$ are close to a common value (otherwise a slightly more cumbersome expression applies).

We extract A from data removing self-citations, which introduce a background of same-gender preference not due to an actual gender preference. This removal is done exactly, as we have a list of all references where all authors are identified with a unique code. The removal of self-citations reduces same-gender citations, introducing a small gender discrimination of order

[6]The sub-sample of "ambiguous" authors (whose name is associated to different genders in different countries) does not show anomalous features that would support the hypothesis of a collective gender bias.

$1/N_G^{\rm aut}$ in the asymmetry: when considering many authors this bias is negligibly smaller than the statistical uncertainty of A, which scales as $1/\sqrt{N_G^{\rm aut}}$.[7] Fig. 3 shows the time evolution of the gender asymmetry, found to be compatible with zero at all times.[8] Restricting to papers after 2010 we find the result shown in table 1.[9] The uncertainty is shown as one standard deviation after the $\pm$ symbol. A hint of an asymmetry, $A_{\rm other} = (4.8 \pm 1.2)\%$, is observed among other about 10^4 papers (mostly unpublished) not included in the 8 major arXiv categories relevant for fundamental physics. As a result, combining all single-author papers citing single-author papers, gives an asymmetry $A_{\rm published} = (1.0 \pm 0.5)\%$ when restricting to published papers, or $A_{\rm all} = (1.9 \pm 0.4)\%$ when including all papers.

The definition of the gender asymmetry could be extended to multi-authored papers knowing how a hypothetical gender bias would depend on the relative amount of F and M authors. One simple possibility is just generalizing the definition of $N^{\rm cit}_{G\to G'}$ into $\sum_{\rm citations} f_G f'_{G'}$ where f_G (f'_G) is the fraction of authors with gender G in each citing (cited) paper. We drop all self citations, now defined as whenever the cited and citing paper have at least one author in common. With the new $N^{\rm cit}_{G\to G'}$ we find the result in table 2. Uncertainties (not shown) are there about 5 times smaller than in the single-author sample, if propagation of errors is naively applied to fractional counts.

Taking into account the definition of the asymmetry A and the relative number of F and M authors in our data, we conclude that A is so close to zero that a non-zero gender asymmetry in citations within its measured range would not significantly distort the bibliometric indices based on citations discussed in the following.

3.2 Hiring

The lack of a gender asymmetry in citations means that there is no fracture along gender lines in the community about which research in fundamental physics is more relevant/used/visible. In section 2.3 we showed that appropriate bibliometric indices based on citations are useful proxies for scientific merit. We here use such indices to search for a possible gender difference in hiring. For each hired or pseudo-hired author we compute his/her bibliometric indices at the hiring moment, defined as in section 2.2. From this we extract the mean bibliometric indices of hired F and M authors.

The left (right) panel of fig. 4 shows the mean number of fractionally-counted papers (of individual citations $N_{\rm icit}$) of authors at their hiring date as reported by InSpire. For the sake of clarity we use traditional color codes: blue (pink) for male (female) authors.

We see that hired F authors do not have, on average, bibliometric indicators above those

[7] More precisely, the uncertainty on A equals $[N^{\rm cit}_{F\to F} N^{\rm cit}_{F\to M}/N^{\rm cit3}_{F\to} + N^{\rm cit}_{M\to F} N^{\rm cit}_{M\to M}/N^{\rm cit3}_{M\to}]^{1/2}$ using the usual propagation of statistical fluctuations on each counts, $\sqrt{N^{\rm cit}_{G\to G'}}$.

[8] An analysis performed along the same lines but replacing genders with countries shows an order one preference for citing authors of the same country, especially in some countries. This can be a manifestation of the stronger contacts between nearby authors.

[9] Our results have been reproduced by [Hossenfelder et al.(2018)], who also try to go beyond the asymmetry by assuming a model similar to our eq. (4) (but with $N_G^{\rm aut}$ replaced by $N_G^{\rm pap}$). As such models introduce questionable systematic uncertainties, we restrict our attention to the model-independent gender asymmetry.

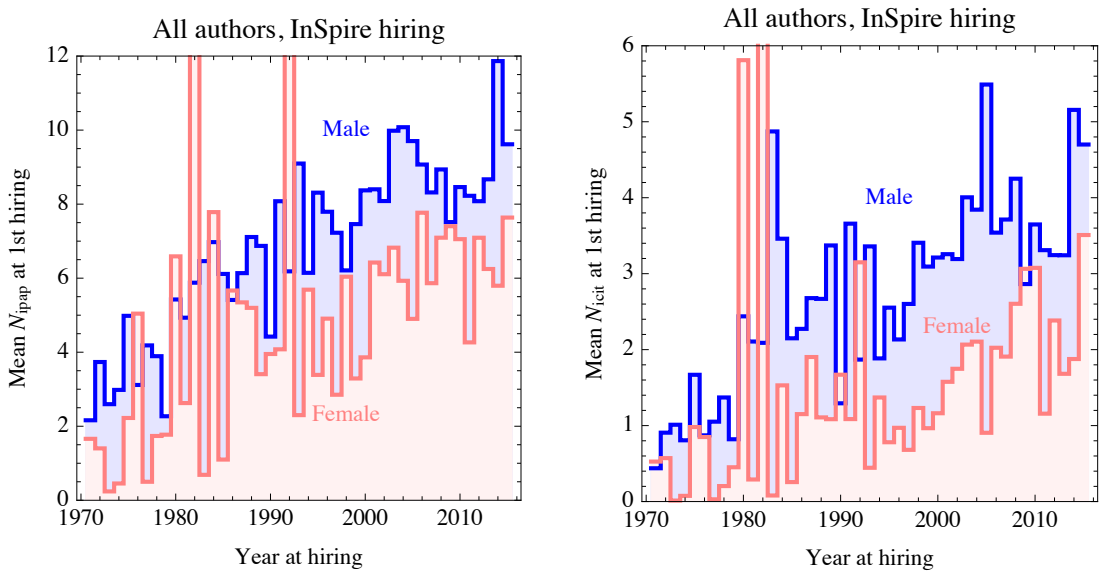

Figure 4: *The left (right) panel shows the mean number of fractionally-counted papers* $N_{\rm ipap}$ *(of individual citations* $N_{\rm icit}$ *) of authors in fundamental physics at the moment of their first hiring, as function of the hiring year. Data are shown separately for male (blue) and female (pink) authors, and compensated for gender history as described in eq.* (2).

of hired M authors. Rather, a tendency in the opposite direction seems present at all times, across the main sub-fields[10] and most countries (statistical uncertainties become significant when restricting to some countries with not enough authors). This result persists after taking into account the possible confounding variables considered in the supplementary section S1.1.

We next provide extra information.

Fig. 5 shows the cumulative distribution of hired physicists as function of their scientific age at hiring. It exhibits no significant gender difference. A gender difference could have been produced in various ways: 1) Some hiring committees might take into account career gaps due to maternity (about which no information is available): this would tend to increase the average scientific age of female hired scientists. 2) A gender discrimination in hiring would tend to reduce the average scientific age at hiring of scientists with the favoured gender. 3) A gender difference in abandonment rates would tend to reduce the average scientific age at hiring of scientists with the higher abandonment rate.[11]

[10] Experimentalists who work in large collaborations tend to have similar bibliometric indicators. The average $N_{\rm icit}$ at hiring can be lower for F authors if they are hired younger than M authors.

[11] The temporal distribution of 245 hires of astronomers in the US after 2010 was studied in [Flaherty (2018)] finding that F authors are hired on average 1.1 ± 0.6 years earlier than M authors (considering the time after receiving the PhD; astronomers are hired on average 5 years later). We find a difference of 0.95 ± 0.5 yr restricting to astro/cosmo authors (considering the time after the first paper; authors are hired on average 9 years later). According to [Flaherty (2018)] the hiring time distribution is better fitted assuming a 3-4 times higher F abandonment rate, rather than assuming a 10:1 bias in favour of F astronomers. However this claim is only based on a very simplified model of hiring that neglects important effects (some authors are better than

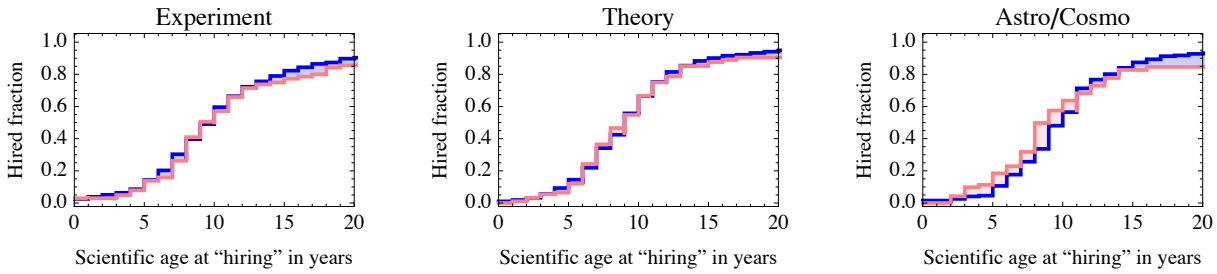

Figure 5: *Among all authors first hired after 2000, we show the cumulative fraction of hired authors as function of their scientific age, for male (blue) and female (pink) authors in experiment (left), theory (middle), astro/cosmo (right). We use* INSPIRE *hires and compensate for gender history as described in eq.* (2).

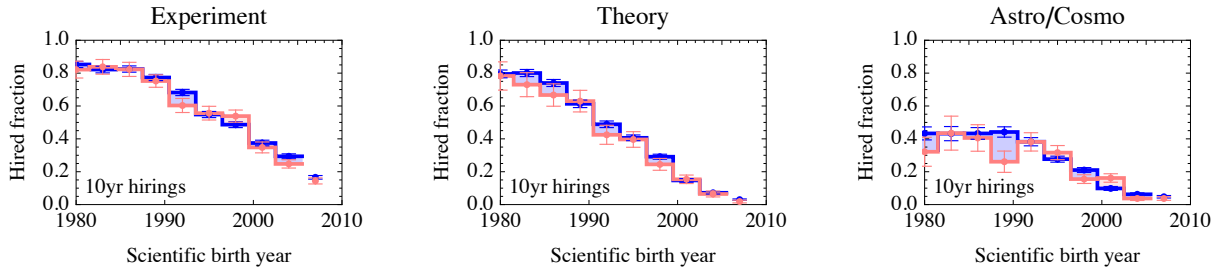

Figure 6: *Fraction of authors hired up to now as function of the date of the first paper. Only the statistical uncertainty is shown; see the text for warnings.*

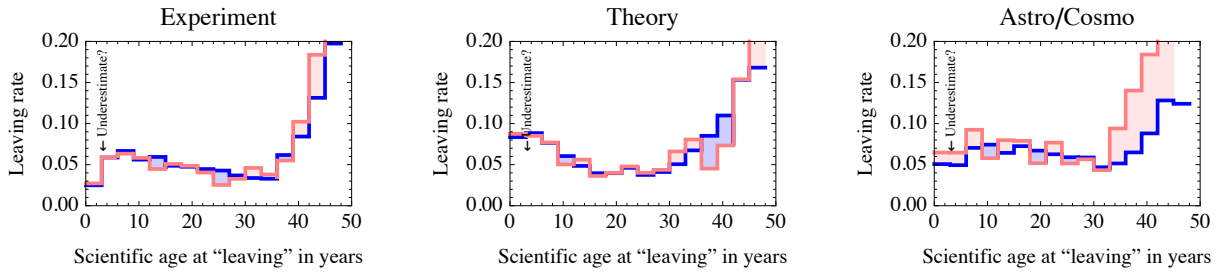

Figure 7: *Fraction of active authors that each year leave research, as function of their scientific age. We considered leavings during 2000-2015, counting as left those authors who no longer wrote a paper up to now (2018) and we compensate for gender history as described in eq.* (2). *See the text for warnings.*

A warning is necessary about the two next plots, which extend the analysis to authors who have not been hired. Our analysis is restricted to InSpire authors listed in HepNames (described in section 2.2), that misses many authors who leave the field after writing a few papers. This generates an extra systematic issue, which presumably tends to be gender-neutral, such that gender ratios presumably are more reliable than absolute rates. Indeed information for M and F authors presumably is similarly incomplete, as InSpire does not collect data about gender, especially of unknown authors.

Fig. 6 shows the fraction of hired authors among those who started writing papers in given time periods. We do not see significant gender differences. We used 10-yr hiring because coverage is here more important than timing. Thereby the plot stops 10 yr ago, and absolute numbers would be different using incomplete InSpire hiring. Furthermore, as warned above, extra un-hired authors not in InSpire would lower the hired fraction.

Fig. 7 shows the abandonment rate per year as function of scientific age. We considered leavings during 2000-2015, counting as left those authors who no longer wrote a paper up to now (2018). Elder authors started when the M fraction was higher and the abandonment rate was lower (as hinted by fig. 6): this confounder generates an apparently lower abandonment rate among M authors. We thereby compensate for gender history as described in eq. (2). We find that the abandonment rate is maximal among elder authors that retire, minimal among senior authors, and intermediate among junior authors (as warned above, we under-estimate the abandonment rate of very young authors that leave the field after writing just a few papers). Abandonment rates show no significant gender difference, in agreement with the null result by [Perley (2019)] and in disagreement with [Flaherty (2018)] (these authors only considered astrophysics).

In conclusion, the gender gap in representation at the entrance level of research is negligibly affected by 'leaky pipeline' effects consistently with [Ceci et al.(2014)] that finds large gender differences at PhD level in STEM, and mild differences in the subsequent progress; see also [Miller et al.(2015), Allen-Hermanson (2017)].

3.3 Productivity

In this section we study scientific productivity as quantified through bibliometric indices. Of course, such indices say nothing about other activities of researchers that do not result in publications, such as teaching, mentoring and outreach. Figures 1, 2, 8, 11 show a possible gender gap in the fractionally-counted number of papers: male authors write, on average, a few 10% more papers. The gap is consistent with earlier findings in the literature (see e.g. table 2 of [Ceci et al.(2014)] and [Abramo et al.(2009)], [Abramo et al.(2015)]). A slightly larger gap is found in the number of received individual citations.

others; quotas would not be overfilled, etc). We do not attempt modelling hiring, as we do not see how models can be made realistic. Rather, we have extra data about papers and citations which support neither a 10:1 bias (see fig. 4) nor a 4:1 difference in abandonment rates (see fig. 7). Results by [Flaherty (2018)] have been "firmly ruled out" by [Perley (2019)], who ruled out gender differences larger than 40% in hiring and abandonment rates.

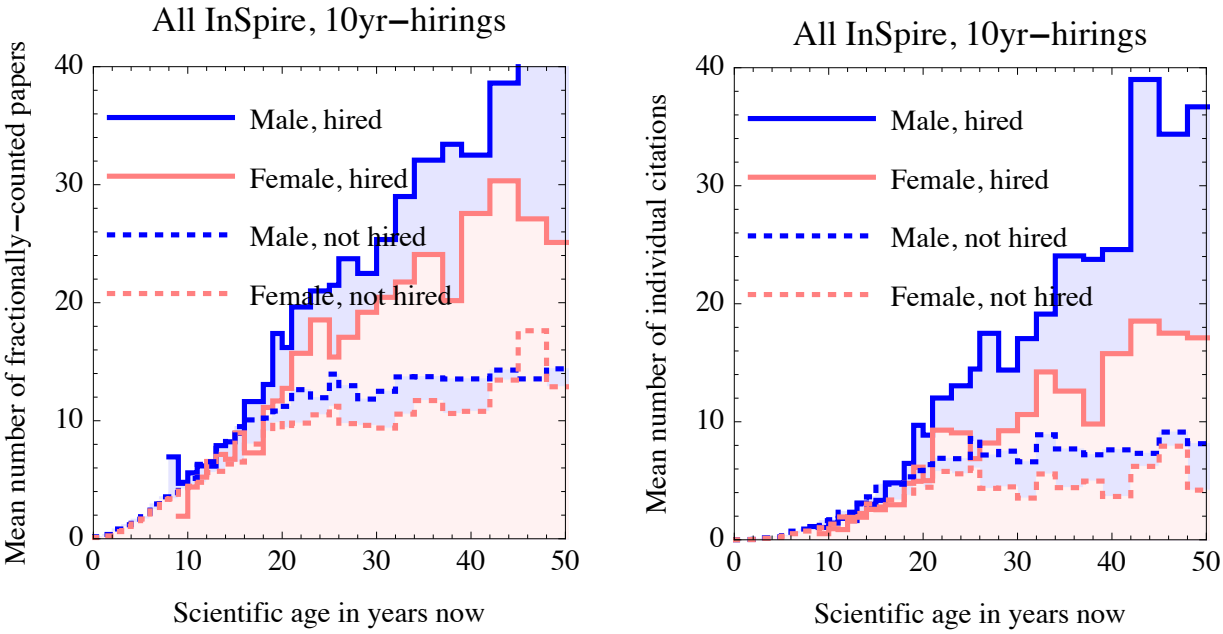

Figure 8: *Mean number of fractionally-counted papers (left) and of individual citations (right) as function of scientific age (time after the first paper) of scientifically active authors now.*

Is such gap due to the different average scientific age of M and F authors? In order to check this possibility, the left panel of fig. 8 shows the mean number of fractionally-counted papers written by M and F authors as function of their scientific age (time since their earliest paper). The gap persists. The right panel of fig. 8 similarly shows the mean number of received individual citations. In both cases we see that junior M and F authors have similar productivity, and that a gap develops with their scientific age. A higher scientific age means going backwards in time, to authors that started earlier when the field was different and when the F percentage was smaller.

The averages in fig. 8 are shown separately for hired and not-hired authors, using 10yr-hires in order to have a more complete coverage. We see that hiring is not the reason of the gap.

Furthermore, in fig. 8 we only considered scientifically active authors (those who wrote at least one paper after 2013), such that these results would not be affected by a gap in abandonment rates

Figure 8 does not compensate for possible career gaps, as such gaps do not exhibit significant gender differences. This is shown in fig. 9, where for each author we computed the longest time gap between consecutive papers, using arXiv dates to have precise information about publication dates. The distribution of longest gaps among M and F authors shown in fig. 9 does not exhibit significant gender differences. A similar null result is found restricting to hired authors. Stopping writing papers might however be the extremum of a tendency towards reduced productivity (possibly due to maternity issues). We thereby searched for consecutive years of reduced publication intensity: some authors are more regular, other experience periods

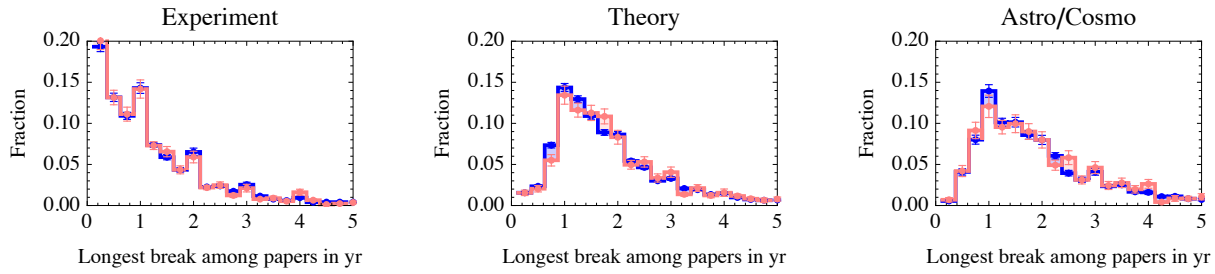

Figure 9: *Fraction of authors active between 2000 and now (divided by their main topic) as function of the longest time break among their papers. We compensate for gender history as described in eq.* (2).

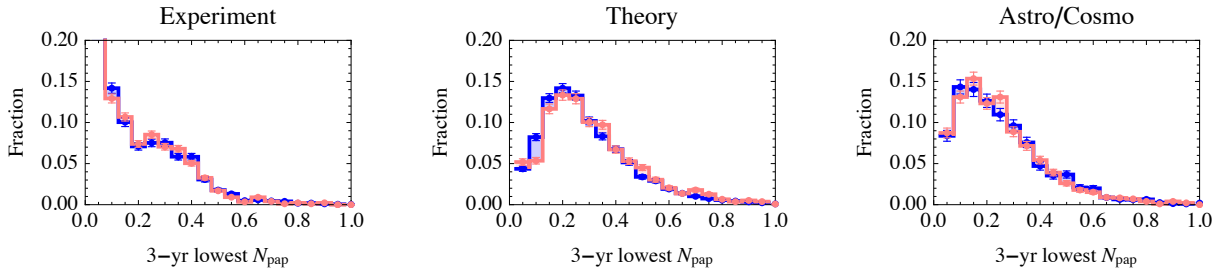

Figure 10: *For each author we compute the minimal number of papers he/she produced in a consecutive 3 year period. We divide this number by the author average publication rate, obtaining a number* $0 \leq r \leq 1$ *normalised such that* $r = 1$ *indicates an author who published in a regular way, while* $r = 0$ *indicates an author with a 3 year period of null productivity. As function of* r *we plot the fraction of authors active between 2000 and now, divided by their main topic. We compensate for gender history as described in eq.* (2).

of relatively lower productivity, but again the distributions show no significant gender differences, see fig. 10. No significant gender differences are found looking at periods of relatively higher productivity.

A gender difference in abandonment rates or career gaps or periods of lower productivity would reduce the cumulative number of papers and of received citations of authors at career level. It is thereby interesting to test if a gap persists in non-cumulative productivity indices that avoid summing over author careers. We proceed as follows: each year we select the subset of scientifically active authors that produced papers, and show in fig. 11 their average productivity, separately for M and F authors. We find that active F authors produce on average roughly 30% less papers than active M authors, and receive roughly half citations.[12]

[12]We adopted fractional counting. Using full counting, hyper-authored publications would lead to a recent

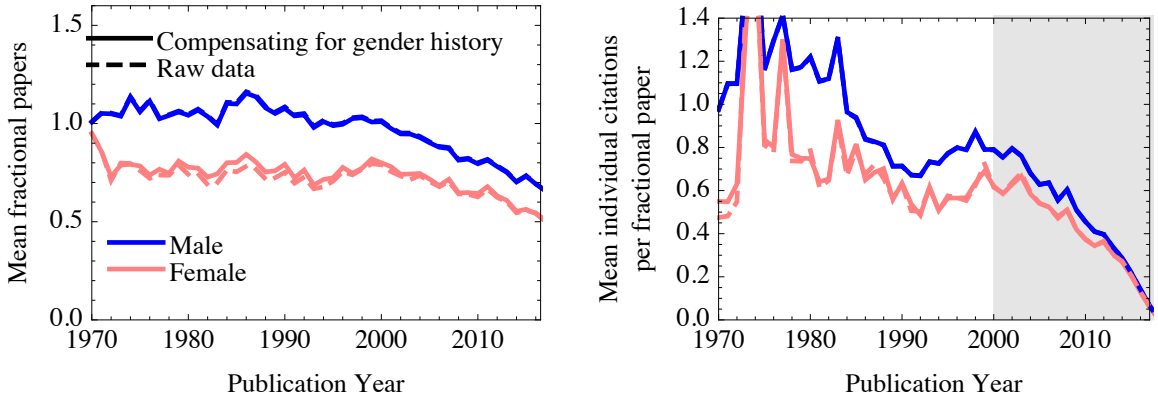

Figure 11: **Left**: *mean number of fractionally-counted papers produced each year by M and F authors active that year.* **Right**: *mean number of received individual citations divided by mean number of fractionally-counted papers. The shading reminds that citation counts are incomplete for recent papers. The continuous curves show the result compensated for gender history as described in eq. (2), and the negligibly different dashed curves show raw data.*

Furthermore, fig. 12 (left panel) analyses the gap at the level of papers, finding a smaller F percentage among authors of top-cited papers, even when restricting to single-author papers (see also [West et al.(2013), Jagsi et al.(2006)]). The right panel of fig. 12 shows that F authors tend to work in larger collaborations.

The supplementary section S1.2 discusses other possible confounding variables, without finding anything that can remove the gender gap in productivity.

We then discuss some possible causes of such gap.

In various countries F authors have earlier retirement ages. But gender differences show up before retirement in fig. 8. Furthermore many physicists tend to remain scientifically active after retirement (although the productivity of most physicists tends to decline before retirement).

A possible reason for the gender gap observed in various fields is children and maternity, see [Ceci et al.(2014)] for a recent summary of the literature, which is not univocal. Some studies find no or small effect [Cole et al.(1987), Sax et al.(2002), Xie et al.(2003), Stack (2004)], other studies find a negative impact (on women [Fox (1995), Ginther et al.(2009)], on men and women equally [Hargens et al.(1978)]), other studies found a positive impact (on men [Ceci et al.(2014)], possibly due to selection effects). Results vary depending on field (with physical sciences sometimes being an outlier, possibly a fluctuation) and are mostly focused on the situation in the U.S. and on the number of produced papers or worked hours. [Ceci et al.(2014)] conclude: "the presence of children cannot explain the overall gender productivity gaps". While maternity

boom, roughly equal for M and F authors, due to the appearance of collaborations with thousands of authors. Alternatively, the productivity gap can be seen using full counting and restricting to theorists.

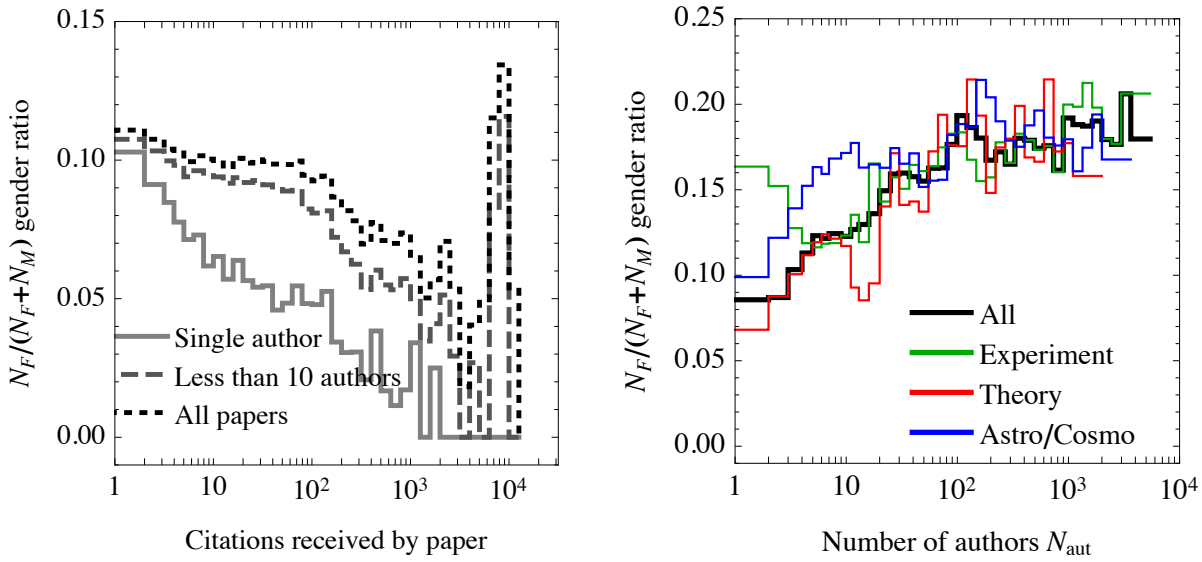

Figure 12: **Left**: *fractional contribution of all F authors as function of the number of individual citations received by the paper. The F fraction of each paper is determined assuming that each author contributed equally to collaboration papers and compensating for gender history. The same result is also computed restricting to solo papers and to papers with less than 10 authors.* **Right**: *fractional contribution of all F authors to papers with $N_{\rm aut}$ authors.*

would deserve a dedicated study, our InSpire data do not provide any personal information so we can only proceed indirectly. As we already described, timing of publications does not show gender differences in periods of null nor of reduced productivity. Fig. 8 indicates that the productivity gap opens at an age roughly consistent with maternity (but also consistent with the transition to scientific independence), and that it does not close at older ages. A similar situation is found analyzing salaries of physicists in the U.S.: no gender gap just after graduation; a 10% gap after 10-15 years according to [Porter et al.(2019)], who report large differences in personal life choices, in particular that women are 4 times more likely to have a career break.

Since maternity laws are different in different countries, an alternative possible strategy is looking for national differences in the M/F gap, which seems stronger in Germany, UK, Italy; weaker in USA, France, null in Japan. However single-country statistics is poor and many other national differences can act as confounding factors.

3.4 Distribution of individual citations

In the previous section we found a productivity gap. We here characterise its statistical properties. Fig. 13 shows the distributions in the number of individual citations $N_{\rm icit}$ received by female and male authors in fundamental physics, considering the whole InSpire data-base. The bell-shaped distributions spread through a few orders of magnitude in $N_{\rm icit}$. The dotted

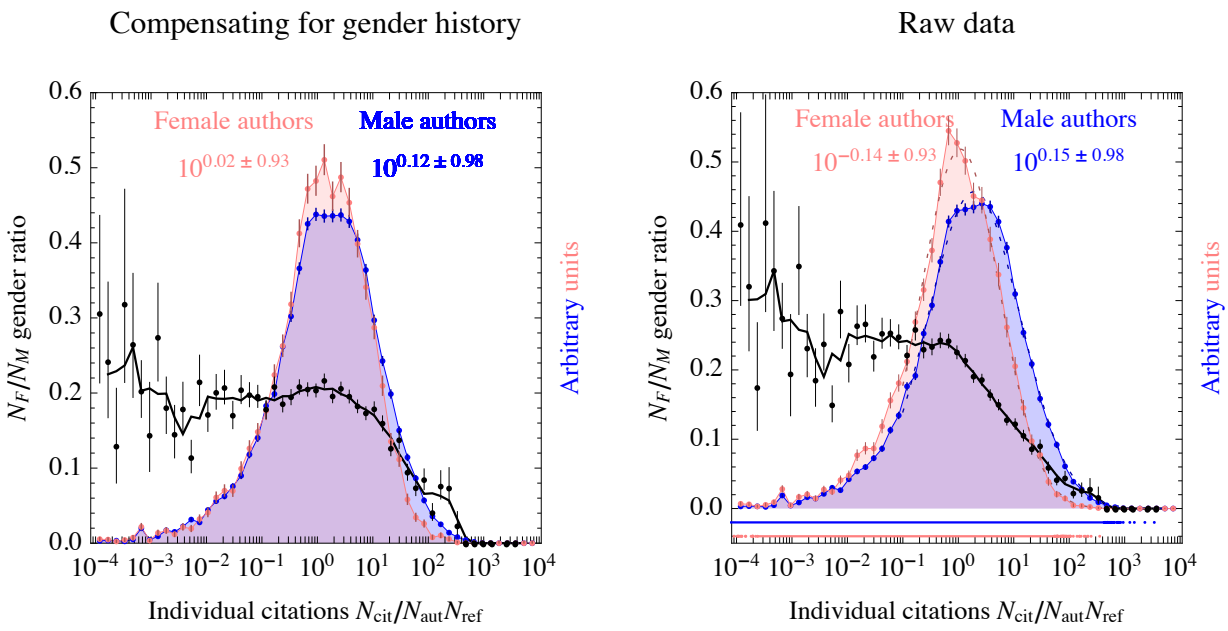

Figure 13: *The bells show the probability distributions of authors in fundamental physics as function of their number of individual citations, separately for Male (blue) and Female (pink) authors and plotted with a common normalization (right axis). The numbers in the figure show the logarithmic average and standard deviation; and the dots below the bells in the right panel show individual authors separately for M and F authors; the dotted curves show how well a log-normal approximates the upper side of the bells. The black smoothened curve (left axis) shows the ratio N_F/N_M between the absolute numbers of female and male authors who received the amount $N_{\rm icit}$ of individual citations indicated on the horizontal axis. Error bars on data points are one standard deviation statistical uncertainties.* **Left***: data after compensating for the higher average age of male authors, as described in eq. (2).* **Right***: raw data, biased by the different age distribution.*

curves show that each bell is well approximated, at least in its upper side, by a log-normal as function of $N_{\rm icit}$ (namely, by a Gaussian as function of $\log N_{\rm icit}$, the variable used on the horizontal axis of fig. 13; log is the logarithm base 10). A log-normal, already observed in bibliometrics [Thelwall et al.(2014)], arises when many positive independent random variables contribute multiplicatively.[13]

The difference between the F and M distributions in fig. 13 is statistically significant. This

[13]The following observations suggest that the dominant random variables are unlikely to be of social type (e.g. the possibility that some authors get more visibility and funds that boost their citation counts [Ruocco et al.(2017)]): we find that the number of citations received by authors in physics tends to grow linearly or quadratically rather than exponentially with their scientific age. Furthermore, looking at single papers, rather than at author careers, a significant excess of young authors (scientific age below ~ 15 years) is observed among authors of top-cited papers.

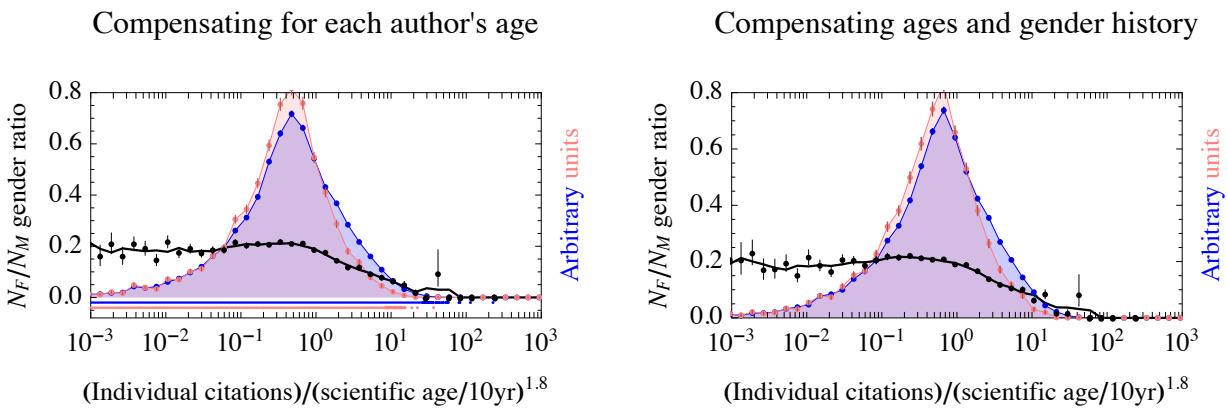

Figure 14: *As in fig. 13, adopting a measure that, on average, does not depend on the scientific age of authors. The M distribution still has a longer upper tail.*

is better shown by the black curve in fig. 13, that shows the ratio N_F/N_M of female versus male authors (left axis) as function of the number of received individual citations (horizontal axis).[14] The N_F/N_M gender ratio is not constant: the M fraction progressively grows when going from average to top authors in terms of individual citations.

We again need to study whether such difference can be a byproduct of confounding factors that affect our composite sample of data. This issue is discussed in the supplementary section S1.3: we don't find any confounder that washes away the trend. As discussed in section 2.4 we compensate for one significant confounder: male authors are presently on average more senior than female authors. This is here relevant because senior authors had more time to receive citations (and because younger authors contribute more to top-cited papers). This confounder does not remove the gender difference in $N_{\rm icit}$, given that it is observed within subsamples of authors with same scientific age (see fig. S11). Correcting for the age confounder is however needed to precisely quantify the difference. We proceed as described in eq. (2). The left panel of fig. 13 shows age-corrected results, while the right panel shows raw data. Both averages and variances differ in raw distributions. The difference in variances persists among the upper sides of the age-corrected distributions, while no significant gender differences are seen in the lower sides of age-corrected distributions, more affected by social phenomena (lower sides would be mostly removed restricting to hired authors).

We can account for different ages using a complementary strategy: we consider a new bibliometric index $N_{\rm icit}/\Delta t_A^p$ that approximatively compensates for the scientific age $\Delta t_A = t_{\rm now} - t_A$ of each author. In order to achieve the compensation we choose $p \approx 1.8$ since $N_{\rm icit}$ averaged over authors approximatively scales as $\Delta t_A^{1.8}$ within all main sub-fields. The result of such correction is shown in the left panel of fig. 14: both M and F distributions become narrower, having removed one source of their variability. We again find a higher male variance among the upper sides. The right panel of fig. 14 shows that applying both age corrections has negligible extra effect. The difference in variances in the upper side is seen in any case.

[14] In precise mathematical notation this is dN_F/dN_M, and the bells are $N_G^{-1} dN_G/d\log N_{\rm cit}$.

The M fraction is largest among top authors, as clear from the dots below the bells in the right panel of fig. 13, which show the individual authors. It is useful to focus on the sub-set of top authors. A physicist might read their names and conclude that no sociological confounder can wash away most of them. It is thereby interesting to show that the gender difference is statistically significant even restricting to top authors. This is done by considering the hypothesis of no gender difference: while being sociologically implausible it has precisely computable consequences, that can be compared to data about top authors. Starting from raw data, the F author with most individual citations is in position $F_1 = 69$. Under the hypothesis of no gender difference one can mathematically compute that the probability of observing $F_1 \geq 69$ is $\wp_1 = m^{F_1-1} \approx 3\ 10^{-6}$, where $m = 1 - f \approx 0.83$ is the male fraction of the large sample (or $m \approx 0.87$ restricting to theorists). Under the same assumption, the k-th F author should be on average position $\langle F_k \rangle = k/f$, with probability distribution $f^k m^{F_k-k}\binom{F_k-1}{k-1}$ [Knapp (2010)]. The observed positions are $F_2 = 147$ (the probability of being in this or lower position is $\wp_2 \approx 3\ 10^{-11}$), $F_3 = 191$ ($\wp_3 \approx 2\ 10^{-13}$), etc., roughly fitted by $F_k \approx 69k$. Such low probabilities mean that gender differences are needed to account for data about top authors. Of course, we already know that the time evolution of $f = N_F/(N_M + N_F)$ is a significant confounder. We again find that such confounder is not enough to remove the difference, since the female fraction f is larger than $1/69$ at all relevant times. More precisely, performing the age correction as in eq. (2), the k-th top female authors shift to $F_k = \{36, 82, 114, 126, \ldots\} \sim 33k$ position. Considering the age-corrected metric $N_{\rm icit}/\Delta t_A^{1.8}$ the positions are $F_k = \{20, 57, 109, 180, \ldots\} \sim 38k$. After correcting for the age confounder $\wp$-values remain small because the female fraction $\sim 1/33$ or $\sim 1/38$ found among top authors remains smaller than the fraction f of female authors in the full sample. An excess of male top-authors is found even after correcting for the age confounder.[15]

The fact that gender differences are maximal among those top authors who receive up to 3000 more individual citations than average authors indicates that differences do not dominantly result from sociological factors that could give factors of 2 differences at individual level (such as harder working vs career gaps, more research vs more teaching, specialisation on physics vs wider interests...).

As the variation of N_F/N_M survives to confounding variables, we try to better understand this effect by investigating its quantitative shape.

A look at fig. 13 or 14 suggests that the M bell has a longer tail of top-authors (raw data show also a difference in averages, mostly due to confounders, that makes the difference in variances less easily visible). The supplementary section S2 shows that the difference in upper variances is statistically significant. We here provide a simpler, more intuitive, argument through analytical approximations based on the observation that the distributions of individual

[15] A varying gender fraction that culminates in a small group of extremely productive, mostly male, 'star authors' has been observed in [Bordons et al.(2003), Abramo et al.(2009)b, Abramo et al.(2015)] (see also [Kwiek (2016)]). [Ioannidis et al.(2019)] used Scopus data about 6.9 million scientists in all disciplines to compute a 'composite' bibliometric indicator, producing a list of top authors. When restricted to fundamental physics, their list (despite minor problematic aspects) is significantly correlated to our list. In their all-fields list the female authors are found in positions $F_k = \{133, 146, 160, \ldots\}$. While this naively seems to extend our findings, we cannot correct their list for the age confounder nor for other confounders.

citations received by M and F authors separately are approximatively log-normal (Gaussian as function of $\ell = \ln N_{\rm icit}$) in their upper side. Log-normal distributions with a common standard deviation σ and different averages $\mu_M \neq \mu_F$ for M and F authors would produce a N_F/N_M of the form

$$\frac{N_F}{N_M} \propto \frac{e^{-(\ell-\mu_F)^2/2\sigma^2}}{e^{-(\ell-\mu_M)^2/2\sigma^2}} = \exp\left[-\frac{R_\sigma}{2}\left(\ell - \frac{\mu_M+\mu_F}{2}\right)\right], \qquad R_\sigma = 2\frac{\mu_M-\mu_F}{\sigma^2}. \tag{5}$$

Different standard deviations $\sigma_M \neq \sigma_F$ would produce

$$\frac{N_F}{N_M} \propto \frac{e^{-(\ell-\mu)^2/2\sigma_F^2}}{e^{-(\ell-\mu)^2/2\sigma_M^2}} = \exp\left[-\frac{R_\sigma}{2}\left(\ell-\mu\right)^2\right], \qquad R_\sigma = \frac{1}{\sigma_F^2} - \frac{1}{\sigma_M^2}. \tag{6}$$

Thereby a dominant difference in averages (standard deviations) would produce a line (a parabola) when N_F/N_M is plotted as function of $N_{\rm icit}$ in log-log scale. Such plot is shown in fig. 15, again performing the usual corrections for the age confounder. The important point is that, in all cases, N_F/N_M exhibits a parabolic shape along the upper sides of the bells, where the log-normal approximation is accurate enough.[16] In conclusions, data exhibit a gender difference in upper variances.

Is this statistically strong preference an artefact of the complexity of the full data-sample? To answer, we repeat the analysis within the independent sub-samples of fig. S11: plotted in log-log scale they independently tend to show a parabolic (rather than linear) trend in N_F/N_M. The dotted curves in fig. S11 show how well each sub-sample can be fitted in terms of $R_\sigma \approx 2$ and $p \approx 2$. Furthermore, the probabilities $\wp_i$ that test the hypothesis of no gender difference restricting to top-authors are small within the sub-samples of fig. S11 (where the top authors are plotted as points), consistently with [Bordons et al.(2003), Abramo et al.(2009)b].

A similar difference in upper variances is also found looking at different bibliometric indicators, see the supplementary section S1.3.

3.5 Self-references

Gender differences in self-references are an interesting topic on its own, as as well as a possible confounding factor to previous analyses. We verified that our previous results persist dropping self-references, and we next justify why there is no need of dropping self-references.

Bibliometric studies found that men cite their own papers more than women: [Cameron et al.(2016)] focused on six ecology journals; [Ghiasi et al.(2006)] on the Web of Science database; [King et al.(2017)] on the JSTOR database; [Hossenfelder et al.(2018)] on single-authored papers in arXiv. In agreement with such studies, restricting to solo papers, we find a $\sim 20-30\%$ higher fraction of self-references among the papers written by M authors (7.3% versus 5.9%).

However, [King et al.(2017)] mention the possibility that this gender difference in self-references is just a reflection of the fact that male authors tend to write more solo papers, and

[16] Adding higher-order terms in the exponent would not change the above conclusion, because fits to the observed distributions find small higher-order terms.

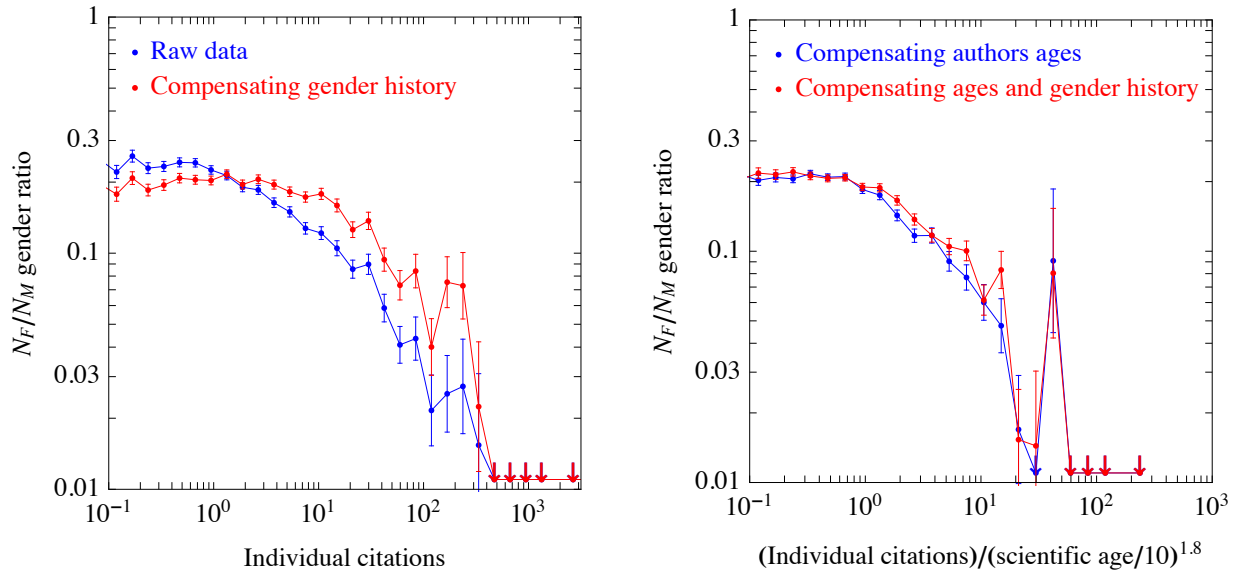

Figure 15: *In the* **left panel** *the points with Gaussian* 1σ *statistical errors are data about the* N_F/N_M *female/male ratio as function of the number of individual citations received. Points with* $N_M > N_F = 0$ *are shown as down arrows. The blue points are raw data, the red points are corrected compensating for the different time evolution of the overall number of* M *and* F *authors. In the* **right panel** *we consider a different bibliometric index which approximatively compensates also for the scientific age of each author. Data are not well fitted by a linear function in log-log scale (which corresponds to* $p = 1$ *in eq. (8)) and can be fitted by a quadratic function (which corresponds to* $p = 2$*). This can be interpreted as different male and female variances, as in eq. (6), rather than as different averages, as in eq. (5).*

thereby have more scientific reasons for self-references. [King et al.(2017)] could not check if this is a significant confounding factor because authors are not disambiguated in their data-base. As we have disambiguated authors, we perform this check finding that this confounding factor removes the gender difference in self-references: a similar fraction of self-references is found comparing male and female authors who wrote the same number of papers $N_{\rm pap}$. In other terms, the fraction of self-references can be a misleading indicator because the average number of self-references grows with the number of solo papers following a scale law $N_{\rm cit}^{\rm self} \propto N_{\rm pap}^{p}$ with power $p > 1$. Our data suggest $p \approx 1.3$.

Since solo papers are a relatively small sub-sample that might be not representative of the full data-base, we extend the analysis to multi-authored papers. In order to do so, we need to distinguish self-references from self-citations. We count a citation as self-citation whenever the citing paper has at least one author in common with the cited paper. We count it as a self-reference only for those authors who wrote both the cited and the citing paper. We clarify with an example. One paper by authors A and B cites a paper by authors B and C: B gives a self-reference; both B and C receive a self-citation (B directly and C indirectly).

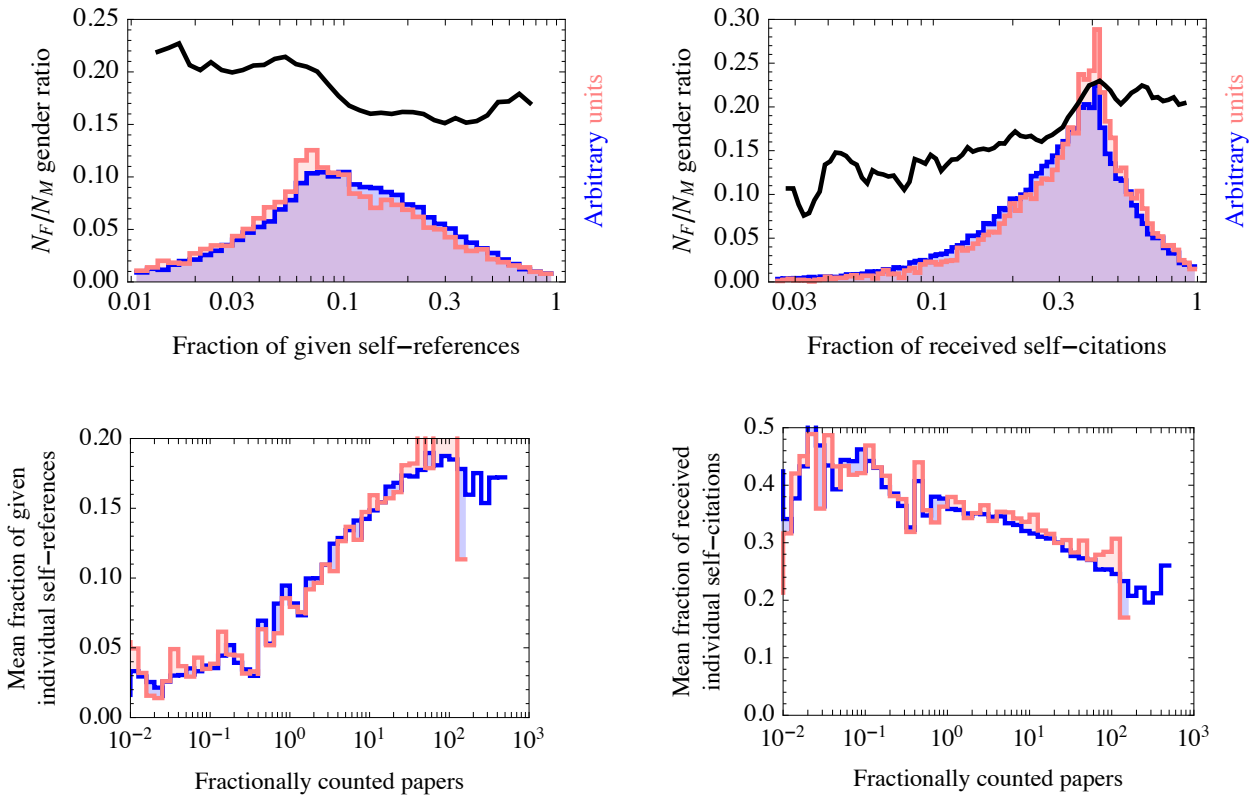

Figure 16: **Left**: *distribution of authors as function of their self-reference rate.* **Right**: *distribution of authors as function of their self-citation rate. Bottom row: mean fraction of given self-references (left) and of received self-citations (right) among M (blue) and F (pink) authors with the number of fractionally-counted number of papers indicated on the horizontal axis.*

As usual, we adopt fractional counting. For each author we compute the total number of received individual citations $N_{\rm icit}$, of received individual self-citations $N_{\rm icit}^{\rm received}$, of given individual self-references $N_{\rm icit}^{\rm given}$, and of fractionally-counted papers $N_{\rm pap}$ (equal to the number of given individual references).

We consider the mean fraction of given individual self-references $N_{\rm icit}^{\rm given}/N_{\rm icit}$ (left panels of fig. 16) and the mean fraction of received individual self-citations $N_{\rm icit}^{\rm received}/N_{\rm icit}$ (right panels of fig. 16). The upper row shows some gender difference: male authors tend to give themselves a higher fraction of self-references (13.2% instead than 12.0%); female authors tend to receive a higher fraction of indirect self-citations (23.5% instead then 19.8%). Such differences again disappear in the lower row of fig. 16, where the self fractions are computed as function of the number $N_{\rm pap}$ of fractionally-counted papers written by each author. Similarly to what found in the solo sample, male authors tend to cite more themselves and their collaborators just because male authors tend to have more past papers. Similar results are found restricting within the

main sub-fields.

4 Conclusions

We performed a bibliometric analysis of gender issues in fundamental physics world-wide from $\sim$ 1970 to now. Bibliometrics gives quantitative data on activities of researchers that result in publications and tells nothing about other possible activities such as teaching, mentoring, outreach, etc. Nevertheless, research is an interesting area where individual talent can be expressed, as confirmed by the large differences between authors found in data. Concerning gender we find the following results:

1. First, we see the well-known initial gender difference in representation: there are roughly 4 males for each female among new authors that appear at PhD-level. The initial female fraction is not positively correlated with the Global Gender Gap Index of the countries,[17] and negligibly evolves in the subsequent career stages (fig. 7).

2. When citing works by others, authors exhibit no or small gender difference: male and female authors have the same average opinion about which research in fundamental physics deserves to be cited (fig. 3).[18] Furthermore, M and F authors give to their own papers a similar fraction of self-references (fig. 16), taking into account that M authors tend to write more papers, especially solo papers (fig. 12).

3. Female authors do not have, at hiring moments, higher average bibliometric indicators based on individual citations or fractionally-counted paper than male authors (fig. 4 and 5).

4. Among authors identified by InSpire HepNames profiles (that misses authors who write very few papers) we do not find a gender difference in hired percentages (fig. 6) nor in abandonment rates (fig. 7), nor in longest breaks among papers (fig. 9), nor in periods of reduced activity relatively to their average (fig. 10).

[17] An anti-correlation known as "gender equity paradox" [Stoet et al.(2018)] has been observed among students, which are less mobile than researchers.

[18] Some authors discuss the possibility that physicists are collectively affected by an unconscious gender bias — a concept that received recent attention in the U.S. following the development of Implicit Association Tests (IAT) claimed to reveal such biases. Even if such tests were scientifically valid (see however [Oswald et al.(2015)] for a meta-review), reading a scientific paper involves different mental processes than those probed by IAT. Some of our results are based on citations: the same conclusions are reached removing from our analysis citations to single-authored research, that would be more affected by a hypothetical collective gender bias. Furthermore, fields that study bias might have their own biases: [Winegard et al.(2018), Stewart-Williams et al.(2019)] found that scientific results exhibiting male-favouring differences are perceived as less credible and more offensive. [Handley et al.(2015)] found that men (especially among STEM faculty) evaluate gender bias research less favourably than women.

The above results are in line with the literature, as summarized in [Ceci et al.(2014)]: "the overall picture is one of gender neutrality"; "no evidence of women having harder time getting tenure"; post-bachelor gender differences in attrition rates are significantly smaller in STEM than in life science, psychology and social science. The literature finds a second gender difference, in productivity: "women on average publish fewer papers than men", "there are no sex differences in citations per article" [Ceci et al.(2014)]. We find:

5. A productivity gap both in the fractionally-counted number of publications and in their citational impact (fig. 8), which does not appear to be dominantly concentrated in specific countries, topics, periods, bibliometric indicators, journals (fig. S1), etc. The gap is also found without integrating over careers (see fig. 11, 12).

6. A gradually increasing male fraction when going from average to top authors in terms of individual citations (or other indices, such as fractionally-counted publications). The quantitative shape of this trend appears dominantly due to a higher variance in the upper side of the M distributions (see fig. 15, eq. (5), eq. (6)), rather than due to a difference in averages.

We verified that our results still hold ignoring hyper-authored papers (possibly affected by guest authorship) or restricting to single-authored papers.

While many social phenomena could produce different averages, producing different variances would need something that specifically disadvantages research by top female authors. Just to make one example of social nature, a gender gap in research productivity could arise if better female authors receive more honours and leadership positions that drive them away from research. However, data also show an excess of young authors among those who produced top-cited papers: the excess is observed both among M and F authors. This suggests extending our considerations from possible sociological issues to possible biological issues.

It is interesting to point out that the gender differences in representation and productivity observed in bibliometric data can be explained at face value (one does not need to assume that confounders make things different from what they seem) relying on the combination of two effects documented in the scientific literature: differences in interests [Su et al.(2009), Diekman et al.(2010), Lippa (2010), Su et al.(2015), Thelwall (2018)b] and in variability [Halpern et al.(2007), Deary et al.(2007), Hyde (2014), Stevens et al.(2017), Wang et al.(2013)].

Greater male interest in things and greater female interest in people is observed consistently across cultures and time and is large ($d \approx 1$ i.e. distributions differ by about one standard deviation): such difference in interests dominantly accounts for the initial difference in representation.[19] Difference in variability accounts for the difference in productivity. This is consistent

[19]Large gender differences along the people/things dimension are observed in occupational choices and in academical fields: such differences reproduce within sub-fields [Thelwall (2018)b]. In particular, the female participation is lower in sub-fields closer to physics even within fields with their own cultures, such as "physical and theoretical chemistry" within chemistry [Thelwall (2018)b]. This suggests that the people/things dimension plays a more relevant than role than the different cultures of different fields.

Furthermore, psychology finds that females value careers with positive societal benefits more than do males: some authors propose that women tend more to opt out of STEM because "women tend to endorse communal

with [O'Dea et al.(2018)], that confirms the difference in variabilities looking at grades, and observes that this difference alone cannot reproduce the representation gap.

The amount of higher male variability suggested by bibliometric data in fundamental physics is at the 10% level, roughly consistent with independent observations of presumably relevant traits. While such psychometrics observations dominantly probe the central, most populated, part of the distributions, we reasonably expect that physicists probe the upper tail[20] and that top-cited physicists reach the far-end upper tail. We estimate to reach about 5 standard deviations above the mean, given that this is the maximal deviation statistically expected from a pool of $\sim 10^9$ persons, in Gaussian approximation.

Dealing with complex systems, any simple interpretation can easily be incomplete, including a hypothetical gender discrimination. In any case, it is interesting that data can be explained without invoking such hypothesis.

We conclude addressing ethic and social values, given that a gender difference in variances is seen by some as offensive, like other ideas originally proposed by Darwin [Hill (2017)] (modestly keeping things in proportions in this comparison). The interpretation in terms of different variabilities implies that we should keep giving gender-neutral equal opportunities to everybody by considering each person based on his/her individual qualities, not as member of a demographic group (gender, nationality or whatever). The refusal to consider population level differences in distributions when trying to understand gaps in representation can lead to discriminations allegedly aimed at establishing equal outcomes (see e.g. [Strumia (2019)] for a more extensive discussion of such issues).

Acknowledgements I thank the referees for their comments; the InSpire team for clarifications; Guy Madison for discussions and suggestions; Riccardo Torre for (among many things) having implemented the WGND name-gender association; Sabine Hossenfelder for having independently replicated the results in fig. 8b and section 3.1 using arXiv data [Hossenfelder et al.(2018)]; more colleagues who prefer not to mentioned.

Competing Interests The author has no competing interests. No funding has been used for this research, performed outside working time.

A Data

This study is based on bibliometric data extracted from the InSpire database around mid 2018. While data are already public from the InSpire web site [InSpire(2010)], we selected the key data and

goals more than men" [Diekman et al.(2010), Evans et al.(2009)]. Indeed [Gibney (2017)] finds that women in UK academia report dedicating 10% less time than men to research and 4% more time to teaching and outreach, and [Guarino et al.(2017)] finds that women in U.S. non-STEM fields do more academic service than men. Concerning fundamental physics, old discoveries gave huge societal benefits, but no practical applications are resulting from contemporary explorations of very small and very large scales (such as production at colliders of particles that decay in 10^{-25} seconds, or cosmological observations of objects billions of light years far away).

[20] Physics attracted students with high average grades, see e.g. fig. 9 of [Ceci et al.(2014)] and fig. 1 of [Ginther et al.(2015)]. Looking at career-integrated citations, physics (and especially fundamental physics) shows the largest ratio between high (90th) and low (25th) percentile (tables 1 and S3 of [Ioannidis et al.(2019)]).

converted them into user-friendly MATHEMATICA format. The code used to extract data is available at GITHUB (link: github.com/RiccardoTorre/InSpireTools) and the resulting data at ZENODO (link: zenodo.org/record/3482884#.XaCNsi2B2L4).

The data consists of a table with more than one million of entries, one for each paper. Each entry contains the following information: 1) an integer number identifying the paper; 2) publication date; 3) publication date on arXiv when available; 4) number of authors; 5) number of references; 6) number of citations; 7) and 8) PageRank with and without self-citations; 9) list of authors with each author identified by an integer number; 10) list of references, with each paper identified by the integer in 1); 11) list of citations; 12) title; 13) arXiv main category when available; 14) 1 if published, 0 otherwise; 15) list of lists of affiliations of each author, with each institute identified by an integer number; 16) arXiv sub category when available; 17) arXiv number when available; 18) journal, with each journal identified by an integer number; 19) collaboration(s), with each collaboration identified by an integer number; 20) number of individual citations. The correspondence between integer numbers and papers, authors, institutes, journals, collaborations is available from INSPIRE (for privacy reasons we avoid making our tables public).

References

[Abramo et al.(2009)] G. Abramo, C.A. D'Angelo, A. Caprasecca, *"Gender differences in research productivity: A bibliometric analysis of the Italian academic system"*, Scientometrics 79 (2009) 517.

[Abramo et al.(2009)b] G. Abramo, C.A. D'Angelo, A. Caprasecca, *"The contribution of star scientists to sex differences in research productivity"*, Scientometrics 81 (2009) 137.

[Abramo et al.(2015)] G. Abramo, T. Cicero, C.A. D'Angelo, *"Should the research performance of scientists be distinguished by gender?"*, Journal of Informetrics 9 (2015) 25 [arXiv:1810.12737].

[Abramo et al.(2018)] G. Abramo, C.A. D'Angelo, F. Rosati, *"Selection committees for academic recruitment: does gender matter?"*, Research Evaluation 24 (2015) 392 [arXiv:1810.13236].

[Allen-Hermanson (2017)] S. Allen-Hermanson, *"Leaky Pipeline Myths: In Search of Gender Effects on the Job Market and Early Career Publishing in Philosophy"*, Front. Psychol. 8 (2017) 953.

[Aycock et al.(2019)] L.M. Aycock et al., *"Sexual harassment reported by undergraduate female physicists"*, Phys. Rev. Phys. Education Research 15 (2019) 010121.

[Barthelemy et al.(2016)] R.S. Barthelemy, M. McCormick, C. Henderson, C, *"Gender discrimination in physics and astronomy: Graduate student experiences of sexism and gender microaggressions"*, Physical Review Physics Education Research 12 (2016) 020119.

[Bohm et al.(2014)] G. Bohm, G. Zech, *"Statistics of weighted Poisson events and its applications"*, Nucl. Instrum. Meth. A748 (2014) 1.

[Birnholtz et al.(2006)] J.P. Birnholtz, *"What does it mean to be an author? The intersection of credit, contribution, and collaboration in science"*, J. Ass. Inf. Sci. Tec. 57 (2006) 1758.

[Bordons et al.(2003)] M. Bordons, F. Morillo, M.T. Fernandez, I. Gomez, *"One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists"*, Scientometrics 57 (2003) 159.

[Bornmann et al.(2008)] L. Bornmann, H.D. Daniel, *"What do citation counts measure? A review of studies on citing behavior"*, Journal of Documentation 64 (2008) 45.

[Borsuk et al.(2009)] R. Borsuk et al., *"To name or not to name: The effect of changing author gender on peer review"*, BioScience 59 (2009) 985.

[Breda et al.(2019)] T. Breda, C. Napp, *"Girls comparative advantage in reading can largely explain the gender gap in math-related fields"*, PNAS (2019) 05779.

[Buss et al.(2018)] D.M. Buss, W. von Hippel, *"Psychological Barriers to Evolutionary Psychology: Ideological Bias and Coalitional Adaptations"*, Archives of Scientific Psychology 6 (2018) 148.

[Cameron et al.(2016)] E.Z. Cameron, A.M. White, M.E. Gray, *"Solving the Productivity and Impact Puzzle: Do Men Outperform Women, or Are Metrics Biased?."*, BioScience 66 (2016) biv173

[Caplar et al.(2017)] N. Caplar, S. Tacchella, S. Birrer, *"Quantitative evaluation of gender bias in astronomical publications from citation counts"*, Nature Astronomy 1 (2017) 0141.

[Ceci et al.(2011)] S.J. Ceci, W.M. Williams, *"Understanding current causes of women's underrepresentation in science"*, PNAS 108 (2011) 3157.

[Ceci et al.(2014)] S.J. Ceci, D.K. Ginther, S. Kahn, W.M. Williams, *"Women in academic science: A changing landscape"*, Psychological Science in the Public Interest 15 (2014) 75.

[Ceci et al.(2015)] S.J. Ceci, W.M. Williams, *"Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men"*, Front. Psychol. 6 (2015) 1532.

[Chen et al.(2007)] P. Chen, H. Xie, S. Maslov, and S. Redner, *"Finding scientific gems with google"*, J. Informet. 1 (2007) 8 [arXiv:physics/0604130].

[Cole et al.(1987)] J.R. Cole, H. Zuckerman, *"Marriage, motherhood, and research performance in science"*, Scientific American 256 (1987) 119.

[Deary et al.(2007)] I.J. Deary, P. Irwing, G. Der, T.C. Bates, *"Brother-sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979"*, Intelligence 35 (2007) 451.

[Diekman et al.(2010)] A.B. Diekman, A.M. Johnston, E. K. Clark, *"Seeking congruity between goals and roles: A new look at why women opt out of science, technology, engineering, and mathematics careers"*, Psychological Science 21 (2010) 1051.

[Eaton et al.(2019)] A.A. Eaton, J.F. Saunders, R.K. Jacobson, K. West, *"How Gender and Race Stereotypes Impact the Advancement of Scholars in STEM: Professors Biased Evaluations of Physics and Biology Post-Doctoral Candidates"*, Sex Roles 1 (2019) .

[Edwards et al.(2018)] H.A. Edwards, J. Schroeder, H.L. Dugdale, *"Gender differences in authorships are not associated with publication bias in an evolutionary journal"*, PloS One 29 (2018) e0201725.

[Evans et al.(2009)] C.D. Evans, A.B. Diekman, *"On motivated role selection: Gender beliefs, distant goals, and career interest."*, Psychology of Women Quarterly 33 (2009) 235

[Flaherty (2018)] K. Flaherty, *"The Leaky Pipeline for Postdocs: A study of the time between receiving a PhD and securing a faculty job for male and female astronomers"* [arXiv:1810.01511].

[Fox (1995)] M.F. Fox, *"Women in scientific careers"* (1995), in S. Jasanoff et al. editors, "Handbook of science and technology studies" (pag. 205), Thousand Oaks, Sage.

[Fox (2005)] M.F. Fox, *"Gender, family characteristics, and publication productivity among scientists"*, Social Studies of Science 35 (2005) 131.

[Ghiasi et al.(2006)] G. Ghiasi, V. Lavriere, C.R. Sugimoto, *"Gender Differences in Synchronous and Diachronous Self-citations"*, STI conference (2016) Valencia.

[Gibney (2017)] E. Gibney, *"Teaching load could put female scientists at career disadvantage"*, Nature trend watch (2007) .

[Ginther et al.(2009)] D.K. Ginther, S. Kahn, *"Does science promote women? Evidence from Academia 1973-2001"* in R.B. Freeman et al. (editors), *"Science and engineering careers in the United States (National Bureau of Economic Research Conference Report"* (pag. 163), University of Chicago Press.

[Ginther et al.(2015)] D.K. Ginther, S. Kahn, *"Comment on "Expectations of brilliance underlie gender distributions across academic disciplines""*, Science 24 (2015) 391.

[Glass et al.(2010)] C. Glass, K. Minnotte, *"Recruiting and hiring women in STEM fields"*, J. Divers. High. Educ. 3 (2010) 218.

[Guarino et al.(2017)] C.M. Guarino, V.M.H. Borden, *"Faculty Service Loads and Gender: Are Women Taking Care of the Academic Family?"*, Research in Higher Education 58 (2017) 672.

[Halpern et al.(2007)] D.F. Halpern, C.P. Benbow, D.C. Geary, R.C. Gur, J.S., Hyde, M.A. Gernsbacher, *"The science of sex differences in science and mathematics"*, Psychological Science in the Public Interest 8 (2007) 1.

[Handley et al.(2015)] M. Handley, E.R. Brown, C.A. Moss-Racusin, J.L. Smith, *"Quality of evidence revealing subtle gender biases in science is in the eye of the beholder"*, PNAS 112 (2015) 13201.

[Hargens et al.(1978)] L.L. Hargens, J.C. McCann, B.F. Reskin, *"Productivity and reproductivity: Fertility and professional achievement among research scientists"*, Social Forces 57 (1978) 154.

[Hill (2017)] T.P. Hill, *"An Evolutionary Theory for the Variability Hypothesis"* [arXiv:1703.04184], accepted for publication on the Mathematical Intelligencer but never published see *"Academic Activists Send a Published Paper Down the Memory Hole"*.

[Holman et al.(2018)] L. Holman, D. Stuart-Fox, C.E. Hauser, C. Sugimoto, *"The gender gap in science: How long until women are equally represented?"*, PloS Biol. 16 (2018) e2004956.

[Hooydonk (1997)] G. V. Hooydonk, *"Fractional counting of multiauthored publications: Consequences for the impact of authors"*, JASIS 48 (1997) 10.

[Hossenfelder et al.(2018)] S. Hossenfelder, "*Do women get fewer citations than men?*", talk at Chapman University, 2018/11/29.

[Hyde (2014)] J.S. Hyde, *"Gender Similarities and Differences"*, Ann. Rev. Psy. 65 (2014) 3.1.

[InSpire(2010)] High-Energy Physics Literature INSPIRE Database. A. Holtkamp, S. Mele, T. Simko, and T. Smith (INSPIRE collaboration), *"Inspire: Realizing the dream of a global digital library in high-energy physics"*, 2010. See also "INSPIRE-HEP Documentation" (2018) by the INSPIRE-HEP Collaboration.

[Ioannidis et al.(2019)] J.P.A. Ioannidis, J. Baas, R. Klavans, K.W. Boyack, *"A standardized citation metrics author database annotated for scientific field"*, PLOS biol 17 (2019) e3000384.

[Jagsi et al.(2006)] R. Jagsi, E.A. Guancial, C.C. Worobey, L.E. Henault, Y. Chang et al., *"The gender gap in authorship of academic medical literature — a 35-year perspective"*, N. Engl. J. Med. 355 (2006) 281.

[King et al.(2017)] M.M. King, C.T. Bergstrom, S.J. Correll, J. Jacquet, J.D. West, *"Men Set Their Own Cites High: Gender and Self-Citation Across Fields and Over Time."*, Socius 3 (2017) 1

[Knapp (2010)] M. Knapp, *"Are participation rates sufficient to explain gender differences in chess performance?"*, Proc. Biol. Sci. 277 (2010) 2269.

[Kwiek (2016)] M. Kwiek, *"The European research elite: a cross-national study of highly productive academics in 11 countries"*, Higher Education 71 (2016) 379.

[Lariviere et al.(2013)] V. Lariviere, C. Ni, Y. Gingras, B. Cronin, C.R. Sugimoto, *"Bibliometrics: Global gender disparities in science"*, Nature 504 (2013) 211.

[Levin et al.(1998)] S.G. Levin, P.E. Stephan, *"Gender differences in the rewards to publishing in Academe: Science in the 1970s."*, Sex Roles 38 (1998) 1049

[Leydesdorff et al.(2016)] L. Leydesdorff and H. W. Park, *"Full and fractional counting in bibliometric networks"*, CoRR abs/1611.06943 (2016) [arXiv:1611.06943].

[Ley et al.(2008)] T. Ley, B. Hamilton, *"Gender gap in NIH grant applications"*, Science 322 (2008) 1472.

[Lippa (2010)] R.A. Lippa, *"Gender Differences in Personality and Interests: When, Where, and Why?"*, Social and Personality Psych. Compass 4 (2010) 1098.

[Macaluso et al.(2016)] B. Macaluso, V. Lariviére, T. Sugimoto, C.R. Sugimoto, *"Is Science Built on the Shoulders of Women? A Study of Gender Differences in Contributorship"*, Academic Medicine 91 (2016) 1136.

[Ma et al.(2008)] N. Ma, J. Guan, and Y. Zhao, *"Bringing pagerank to the citation analysis"*, Inf. Process. Manage. 44 (2008) 800.

[Marsh et al.(2011)] H.W. Marsh, U.W. Jayasinghe, N.W. Bond, *"Gender differences in peer reviews of grant applications: A substantive-methodological synergy in support of the null hypothesis model"*, Journal of Informetrics 5 (2011) 167.

[Milkman et al.(2015)] K.L. Milkman, M. Akinola, D. Chugh, *"What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations"*, Journal of Applied Psychology 100 (2015) 1678.

[Miller et al.(2015)] D.I. Miller, J. Wai, *"The bachelor's to Ph.D. STEM pipeline no longer leaks more women than men: a 30-year analysis"*, Front. Psychol. 6 (2015) 37.

[Moldwin et al.(2018)] M.B. Moldwin, M.W. Liemohn, *"High-Citation Papers in Space Physics: Examination of Gender, Country, and Paper Characteristics"*, J. of Geophysical Research 123 (2018) 2557.

[Moss-Racusin et al.(2012)] C.A. Moss-Racusin, J.F. Dovidio, V.L. Brescoll, M.J. Graham, J. Handelsman, *"Science faculty subtle gender biases favor male students"*, PNAS 109 (2012) 41 16474.

[Mutz et al.(2012)] R. Mutz, L. Bornmann, H.D. Daniel, *"Does gender matter in grant peer review? An empirical investigation using the example of the Austrian Science Fund"*, Zeitschrift fur Psychologie 220 (2012) 121.

[NASEM report (2018)] US National Academies of Science, Engineering and Medicine, *"Sexual Harassment of Women"* (2018).

[National Research Council] National Research Council, *"Gender Differences at Critical Transitions in the Careers of Science, Engineering and Mathematics Faculty"*, National Academies Press (2009).

[O'Dea et al.(2018)] R.E. O'Dea, M. Lagisz, M. D. Jennions, S. Nakagawa, *"Gender differences in individual variation in academic grades fail to fit expected patterns for STEM"*, Nature Communications 9 (2018) 3777.

[Oswald et al.(2015)] F.L. Oswald, G. Mitchell, H. Blanton, J. Jaccard, P. Tetlock, *"Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies"*, Journal of Personality and Social Psychology 105 (2013) 171.

[Perianes-Rodiguez et al.(2016)] Perianes-Rodiguez, L. Waltman, N.J. van Eck, *"Constructing bibliometric networks: A comparison between full and fractional counting"*, J. Informetrics 10 (2016) 1178 [arXiv:1607.02452].

[Perley (2019)] D.A. Perley, *"Gender and the Career Outcomes of PhD Astronomers in the United States"* [arXiv:1903.08195].

[Pinski et al.(1976)] G. Pinski, F. Narin, *"Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics"*, Inf. Process. Manag. 12 (1976) 297.

[Porter et al.(2019)] A.M. Porter, R. Ivie, *"Women in Physics and Astronomy"*, American Institute of Physics report, 2019.

[Radicchi et al.(2009)] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, *"Diffusion of scientific credits and the ranking of scientists"*, Physical Review E80 (2009) 056103 [arXiv:0907.1050].

[Rasmussen et al.(2019)] K.C. Rasmussen (she/they) et al., *"The Nonbinary Fraction: Looking Towards the Future of Gender Equity in Astronomy"* [arXiv:1907.04893].

[Rossi et al.(2019)] P. Rossi, A. Strumia, A. Torre, *"Bibliometrics for collaboration works"* [arXiv:1902.01693], to appear in the proceedings of the ISSI 2019 conference.

[Ruocco et al.(2017)] G. Ruocco, C. Daraio, V. Folli, M. Leonetti, *"Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent"*, Palgrave Communications 3 (2017) 17064.

[Sax et al.(2002)] L.J. Sax, L.S. Hagedorn, M. Arredondo, M., F.A. Dicrisi, *"Faculty research productivity: Exploring the role of gender and family-related factors"*, Research in Higher Education 43 (2002) 423.

[Sinatra et al.(2005)] R. Sinatra, P. Deville, M. Szell, D. Wang, A.L. Barabasi, *"A century of physics"*, Nature Physics 11 (2005) 791.

[Stack (2004)] S. Stack, *"Gender, Children and Research Productivity"*, Research in Higher Education 45 (2004) 891.

[Stevens et al.(2017)] S. Stevens, J. Haidt, *"The Google Memo: What Does the Research Say About Gender Differences?"*, Heterodox Academy (2017).

[Stoet et al.(2018)] G. Stoet, D.C. Geary, *"The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education"*, Psychological Science 29 (2018) 581.

[Strumia et al.(2006)] A. Strumia, F. Vissani, *"Neutrino masses and mixings and..."* [arXiv:hep-ph/0606054].

[Strumia et al.(2018)] A. Strumia, R. Torre, *"Bibliometric ranking fundamental physics"*, J. of Informetrics 13 (2019) 515 [arXiv:1803.10713].

[Strumia (2019)] A. Strumia, "*Why Are Women Under-Represented in Physics?*, Quillette (2019).

[Stewart-Williams et al.(2019)] S. Stewart-Williams, A.G. Thomas, J.D. Blackburn C.Y.M. Chan, *"Reactions to Male-Favoring vs. Female-Favoring Sex Differences: A Preregistered Experiment"*, PsyArXiv preprint (2019) nhvsr.

[Su et al.(2009)] R. Su, J. Rounds, P.I. Armstrong, *"Men and things, women and people: A meta-analysis of sex differences in interests"*, Psychol. Bull. 135 (2009) 859.

[Su et al.(2015)] R. Su, J. Rounds, *"All STEM fields are not created equal: People and things interests explain gender disparities across STEM fields"*, Front Psychol. 6 (2015) 189.

[Tahamtan et al.(2019)] I. Tahamtan, L. Bornmann, *"What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published between 2006 and 2018"* [arXiv:1906.04588].

[Thelwall (2018)] M. Thelwall, *"Do females create higher impact research? Scopus citations and Mendeley readers for articles from five countries"*, Journal of Informetrics 12 (2018) 1031 [arXiv:1808.03296].

[Thelwall (2018)b] M. Thelwall, C. Bailey, C. Tobin, N.-A. Bradshaw, *"Gender differences in research areas, methods and topics: Can people and thing orientations explain the results?"*, Journal of Informetrics 12 (2019) 149 [arXiv:1809.01255].

[Thelwall et al.(2014)] M. Thelwall, P. Wilson, *"Regression for citation data: an evaluation of different methods"*, Journal of Infometrics 8 (2014) 963.

[Torvik et al.(2016)] V.I. Torvik, S. Agarwal, *"Ethnea – an instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database*", International Symposium on Science of Science March 22-23, 2016. Ethnea web site.

[Waltman (2015)] L. Waltman, *"A review of the literature on citation impact indicators"*, Journal of Informetrics 10 (2016) 265.

[Wang et al.(2013)] M.T. Wang, J. Eccles, S. Kenney, *"Not lack of ability but more choice: Individual and gender differences in choice of careers in science, technology, engineering, and mathematics"*, Psychological Science 24 (2013) 770.

[Way et al.(2016)] S.F. Way, D.B. Larremore, A. Clauset, *"Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks"*, Proc. 2016 World Wide Web Conference (WWW) (2016) 1169 [arXiv:1602.00795].

[Wenneras et al.(1997)] C. Wenneras, A. Wold, *"Nepotism and sexism in peer-review"*, Nature 387 (1997) 341.

[West et al.(2013)] J.D. West, J. Jacquet, M.M. King, S.J. Correll, C.T. Bergstrom, *"The role of gender in scholarly authorship"*, PloS One 8 (2013) e66212.

[West et al.(2014)] J.D. West, M.C. Jensen, R.J. Dandrea, G.J. Gordon, C.T. Bergstrom, *"Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community"*, JASIST 64 (2014) 787.

[WGND] Worldwide Gender-Name Dictionary.

[Williams et al.(2017)] W.M. Williams, S.J. Ceci, *"2:1 faculty preference for women on STEM tenure track"*, Proceedings of the National Academy of Sciences 112 (2017) 5360.

[Winegard et al.(2018)] B. Winegard, C.J. Clark, C. Hasty, R. Baumeister, *"Equalitarianism: A Source of Liberal Bias"*, SSRN Electronic Journal (2018) 3175680.

[Witteman et al.(2019)] H.O. Witteman, M. Hendricks, S. Straus, C. Tannenbaum, *"Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist"*, The Lancet 393 (2019) 531.

[Wolfinger et al.(2008)] N.H. Wolfinger, M.A. Mason, M. Goulden, *"Problems in the pipeline: Gender, marriage, and fertility in the ivory tower."*, J. Higher Educ. 79 (2008) 388

[Wolfram et al.] Wolfram Mathematica, Determine the Gender of Given Name.

[World Economic Forum] World Economic Forum, *"The Global Gender Gap Report 2016"*.

[Xie et al.(1998)] Y. Xie, K.A. Shauman, *"Sex differences in research productivity: New evidence about an old puzzle"*, American Sociological Review 63 (1998) 847.

[Xie et al.(2003)] Y. Xie, Y. K. Shauman , *"Women in science: Career processes and outcomes"*, Cambridge, Harvard University Press.

[Zitt et al.(2008)] M. Zitt, H. Small, *"Modifying the journal impact factor by fractional citation weighting: The audience factor"*, Journal of the American Society for Information Science and Technology 59 (2008) 1856.

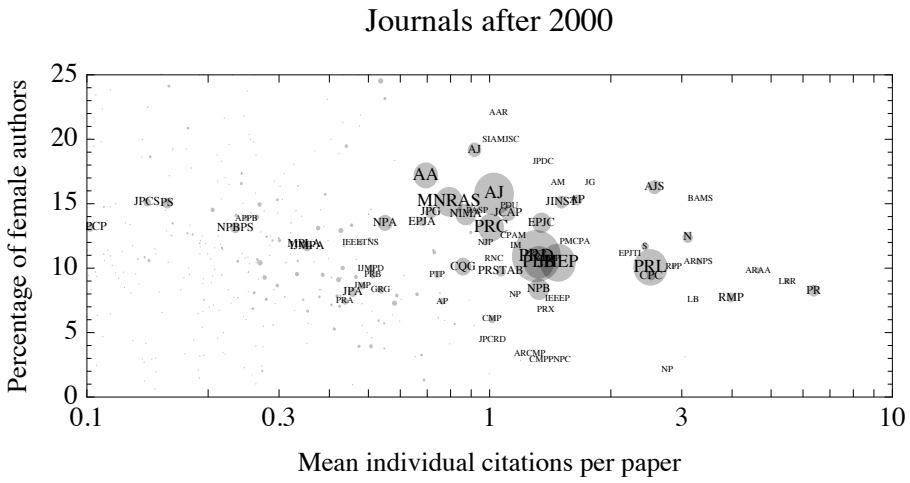

Figure S1: *Journals, classified by the percentage of fractionally-counted female authors and by the average individual citations received by their published papers. Dot areas are proportional to the total bibliometric impact.*

S1 Checking possible confounders

We here show that the gender differences discussed in the main text remain after accounting for possible confounder 'backgrounds'. To start we consider a generic confounder, and in the following sections we consider confounders specific to the various analyses.

While our analyses are not restricted to published papers (because the current main publishing tool in fundamental physics is the pre-print bulletin arXiv) we consider a possible gender bias in the publishing system. Fig. S1 shows that the gender composition among authors of main journals follows the gender composition of their topics (e.g. more F authors in astrophysics) up to small variations. This is consistent with other studies that found gender-fair refereeing processes [Borsuk et al.(2009), Ceci et al.(2014), Edwards et al.(2018)].

S1.1 Hiring

We here study whether confounders can produce the gender difference in the average number of individual citation at hiring moments found in fig. 4 of section 3.2. We explore the following possibilities:

- The fraction of female authors is mildly higher in countries with lower hiring requirements. The top-left panel of fig. S2 shows that the gender difference persists restricting to hires in EU and US, which have similar gender fractions among authors in fundamental physics (see fig. 1).

- The middle-left panel of fig. S2 shows the averages restricted to hiring in top institutions, where research weights more than teaching [Ceci et al.(2014)]. We consider the 100 institutions with the highest number of individual citations received by all their present affiliates (see table 12 of [Strumia et al.(2018)]). Similar results are found using different criteria to select top institutes.

- The top-right panel of fig. S2 shows averages computed separately within the main sub-fields of fundamental physics: similar gender patterns are found.

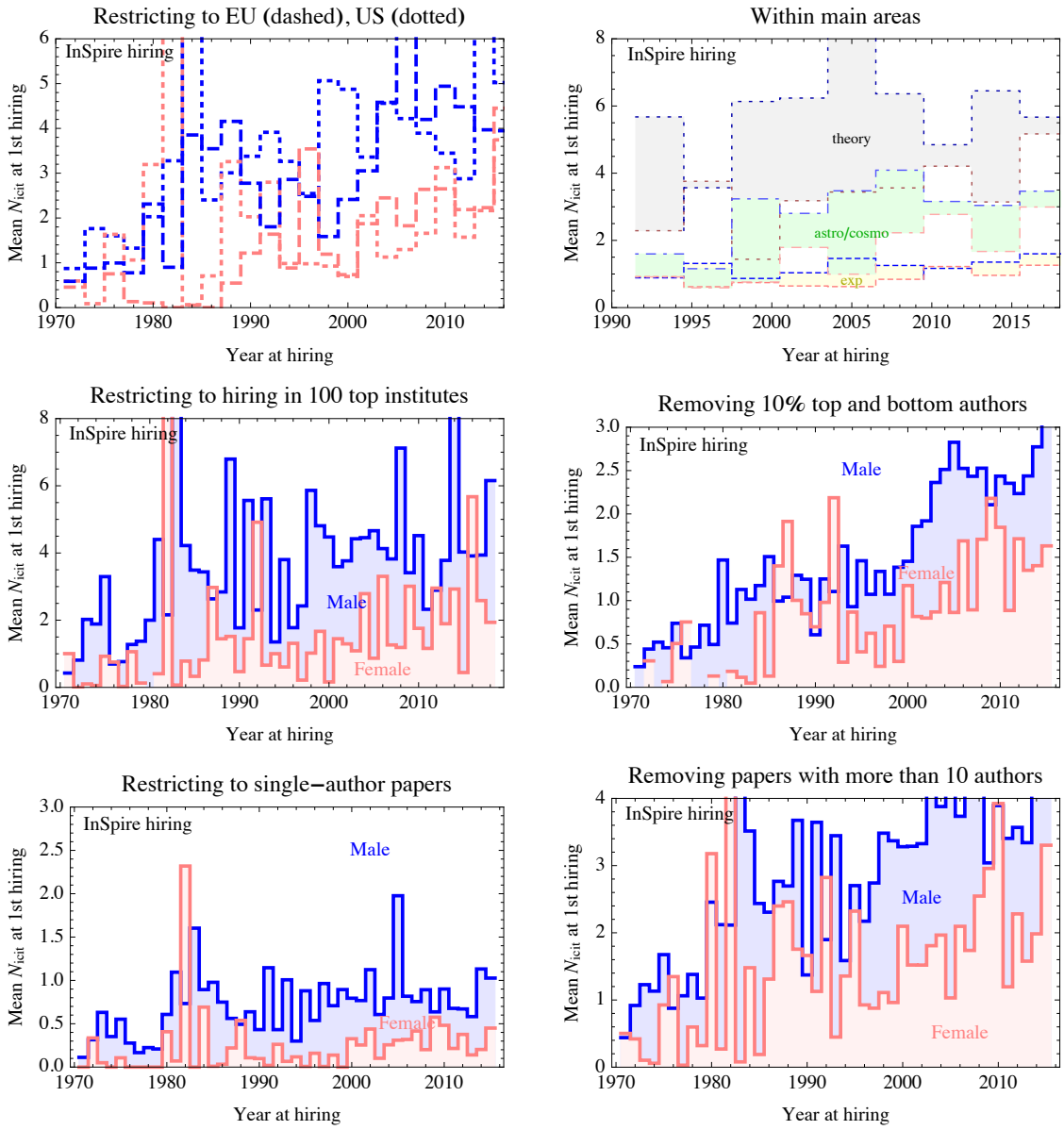

Figure S2: **Top-left**: *as in the left panel of fig. 4, restricting to hiring in EU and US institutions.* **Top-right**: *computing means separately among authors with most papers in theory (dotted), astro/cosmo (dot-dashed), experiment (dashed).* **Middle-left**: *considering only authors hired by 100 top institutes.* **Middle-right**: *recomputing the mean after dropping every year the* 10% *of hired authors with most and with least individual citations at hiring.* **Bottom-left**: *considering only single-author papers.* **Bottom-right**: *considering only papers with less than 10 authors.*

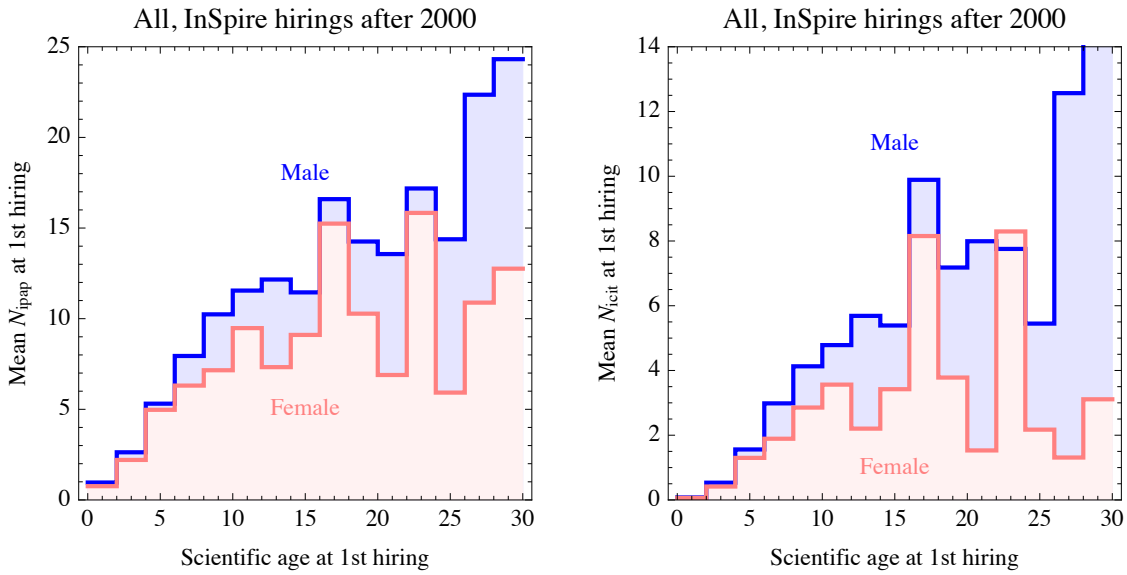

Figure S3: *As in fig. 4, as function of scientific age at hiring (years passed since earliest paper).*

- Rare outliers could affect the mean number of individual citations at hiring moments. The top-right panel of fig. S2 shows the result obtained after removing the top and the bottom 10% of authors. This affects M and F averages in similar ways.
- The bottom panels of fig. S2 show averages computed restricting to solo papers or dropping hyper-authored papers, possibly affected by gift authorship.

Furthermore, qualitatively similar results are found using 5yr or 10yr-hiring; including successive hires and/or removing first hires; using different bibliometric indices such as fractional counting or the 'citation coin' [Strumia et al.(2018)] proposed to discount circular citations including self-citations.

Finally, fig. S3 shows the bibliometric indicators at hiring moments as a function of scientific age at hiring (defined as the time passed since the earliest paper of each author): the gender difference shows no notable dependence on this variable.

S1.2 Productivity

We here study whether the productivity gap found in fig. 8 could be due to confounders.

Checks similar to those already shown in fig. S2 are performed in fig. S4. We restrict to institutions in the EU and US; to main topics; to 100 top institutes; to papers with few authors; we drop top and bottom authors. We find a productivity gap among theorists and astrophysicists but not among experimentalists, who work in large experimental collaborations where bibliometric indicators cannot recognise individual merit (indices of experimentalists are significantly correlated). The productivity gap is found even using non-fractional citation counts (namely, assigning a paper with $N_{\rm aut}$ authors fully to each author) provided that the analysis is restricted to theorists and/or to papers with less than about 10 authors; otherwise average citation counts are dominated by large collaborations.[21]

[21] We thank Sabine Hossenfelder for discussions.

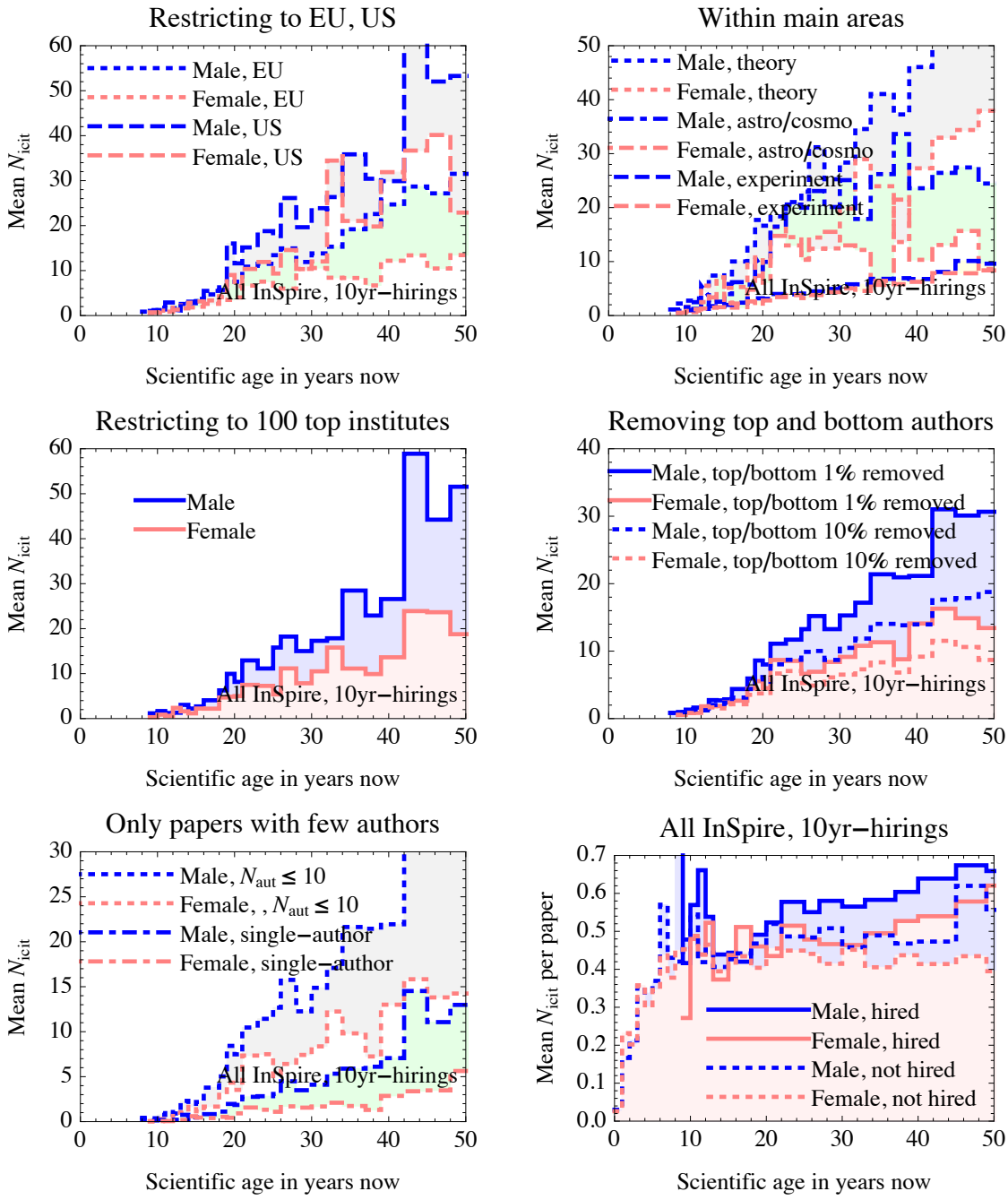

Figure S4: **Top-left**: *as in fig. 8, restricting to authors in EU and US institutions.* **Top-right**: *computing means separately among authors with most papers in theory (dotted), astro/cosmo (dot-dashed).* **Middle-left**: *considering only authors mainly affiliated to 100 top institutes.* **Middle-right**: *recomputing the mean after dropping the* 1% *or* 10% *fraction of hired authors with most and least individual citations.* **Bottom-left**: *considering only papers with few authors.* **Bottom-right**: *considering the average number of fractionally counted citations per fractionally-counted paper.*

Furthermore, a gap in the same direction is observed computing the median, restricting to authors with more than 5 papers; removing solo papers (possibly affected by a hypothetical collective gender bias); replacing individual citations with fractional counting or PageRank-based metrics.

S1.3 Distribution of individual citations

In fig. 13 of section 3.4 we found a higher male fraction among top-authors. We here study whether such trend found can be attributed confounders. The higher average seniority of M authors is a significant confounder, already taken into account in the main text.

One can worry that papers with many authors might contain gift authorships, or that the trend might be a byproduct of gender differences in co-authorship. The trend becomes stronger restricting to papers with less than 10 authors, as well as restricting to single-author papers, see fig. S5. With these cuts, the trend is visible also considering simple citation counting, see fig. S6. Furthermore, fig. 12 shows the percentage female contribution to papers as function of the number of citations received by papers: we see that F authors contributed to high-impact collaboration papers, but not yet to high-impact single-author papers.

One can worry that the trend is due to our specific choice of bibliometric index, individual citations. We thereby repeat the same analysis using different bibliometric indicators:

- Fig. S7 shows the result using the CitationCoin [Strumia et al.(2018)]: we again find a gender difference among top authors, that persists after compensating for the different gender history.
- Fig. S8 considers the mean individual citations per paper received by each author, $N_{\rm icit}/N_{\rm pap}$. Like the CitationCoin, this metric combines the information on the number of citations and of papers. However, it is a less effective figure of merit, as it leads to lists of top authors that contain authors who abandoned the field after writing one or few well-cited papers.[22]
- Fig. S9 shows the result using the AuthorRank computed as in [Strumia et al.(2018)] and shifted such that its lowest value is 0. The trend is again visible.
- Fig. S10 shows that the trend is again present considering the fractionally-counted number of papers written by each author, even dividing by his/her scientific age. The number of papers (published or not) written by each author is a bibliometric indicator less correlated with scientific merit than the number of received individual citations, but more objective (for example, one might worry that citations are affected by a gender bias).

We next check that not only the trend, but also its specific quantitative shape discussed around eq. (8) is not due to confounders. Fig. 15 shows that this is the case when correcting for different author ages. Fig. S11 shows that the trend is present in independent slices of the data-base selected according to possible confounders:

- The top-left panel of fig. S11 shows that the trend is independently present when the data-base is subdivided into the main topics of fundamental physics (experiment, theory and astrophysics, as deduced from arXiv categories available after $\sim$1995). A hint of the trend is present among experimentalists, despite their large recent collaborations.

[22] This metric has 0.36 correlation with $N_{\rm icit}$. The related metric $N_{\rm icit}/N_{\rm ipap}$ (mean individual citations per fractionally-counted paper, 0.04 correlation with $N_{\rm icit}$) magnifies the problem up to the point that top authors are unknown authors who wrote one or few well-cited papers in collaborations [Strumia et al.(2018)].

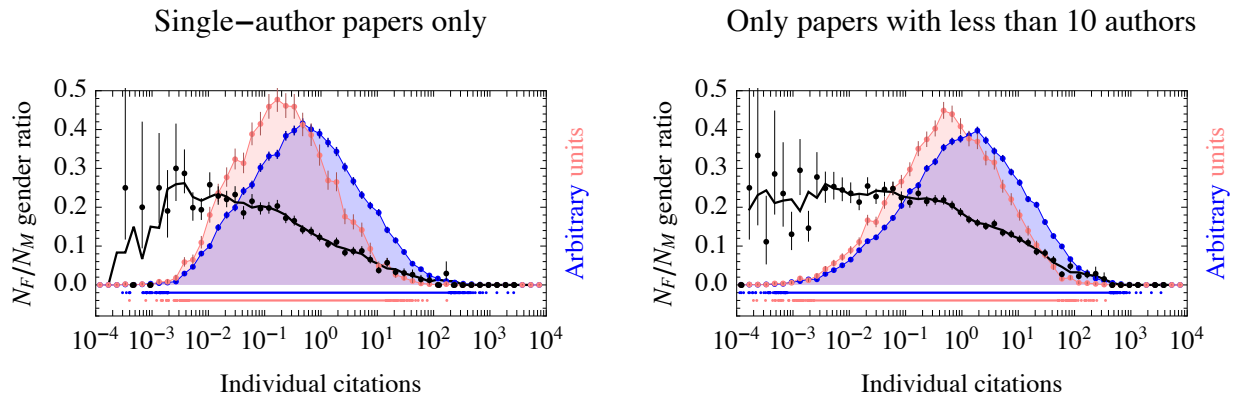

Figure S5: *As in fig. 13, but restricting to single-author papers (left) or to papers with less than 10 authors (right).*

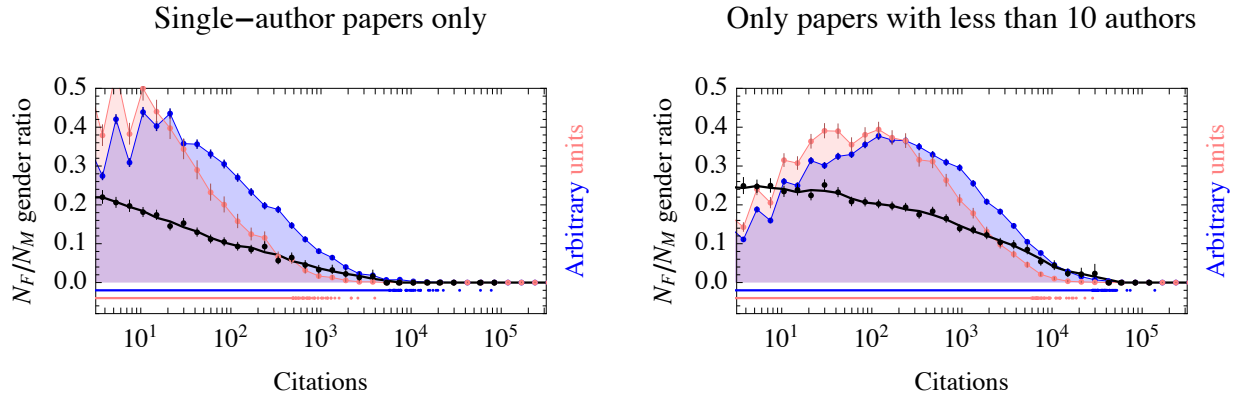

Figure S6: *As in fig. 13, but considering citations.*

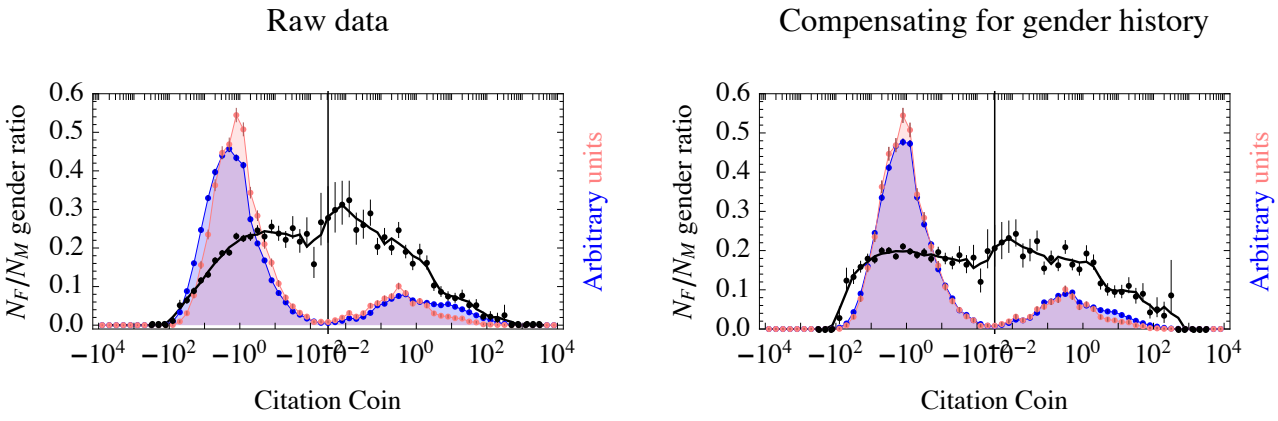

Figure S7: *As in fig. 13, but considering the Citation Coin.*

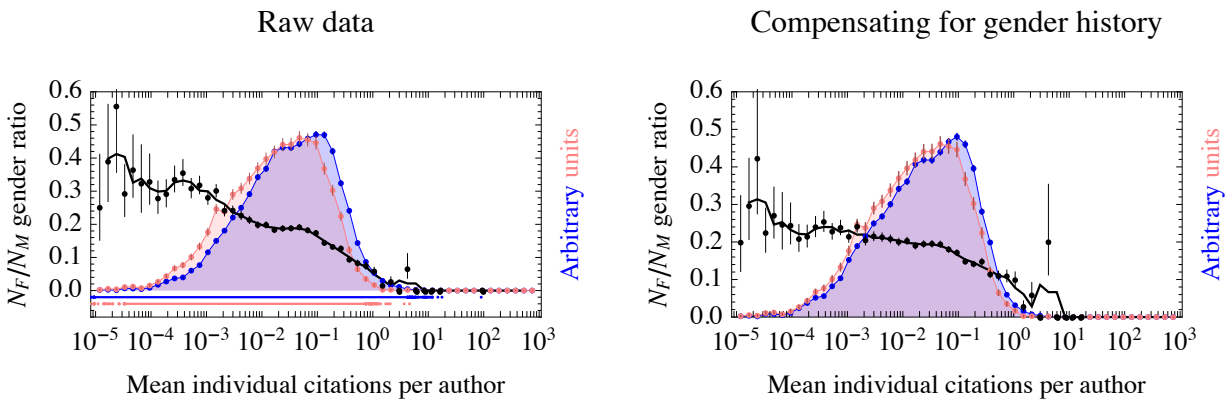

Figure S8: *As in fig. 13, but considering the average number of individual citations received by each author,* $N_{\mathrm{icit}}/N_{\mathrm{pap}}$.

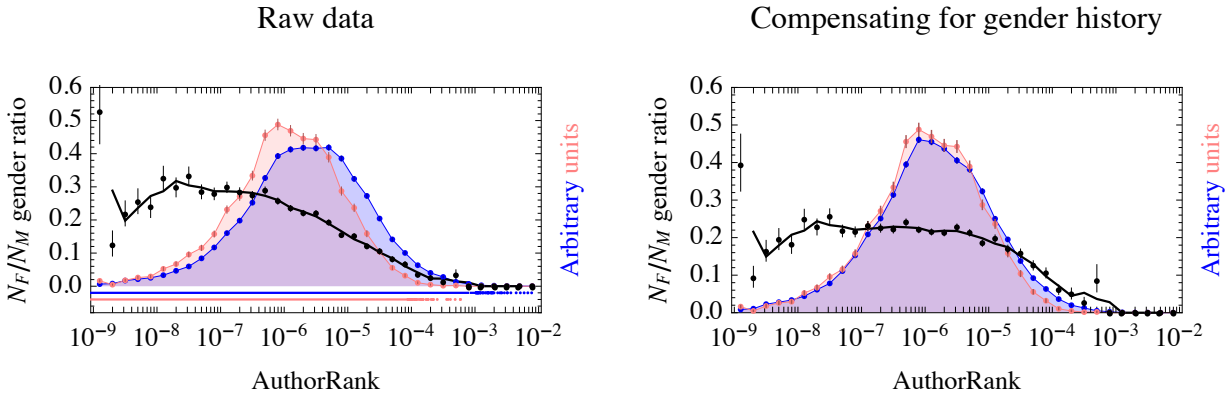

Figure S9: *As in fig. 13, but considering the Author Rank.*

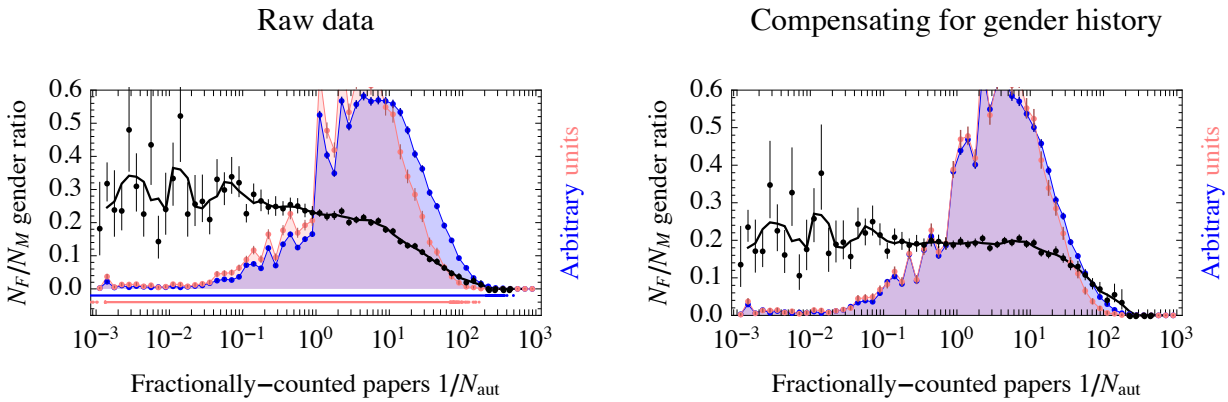

Figure S10: *As in fig. 13, but considering the number of fractionally-counted papers.*

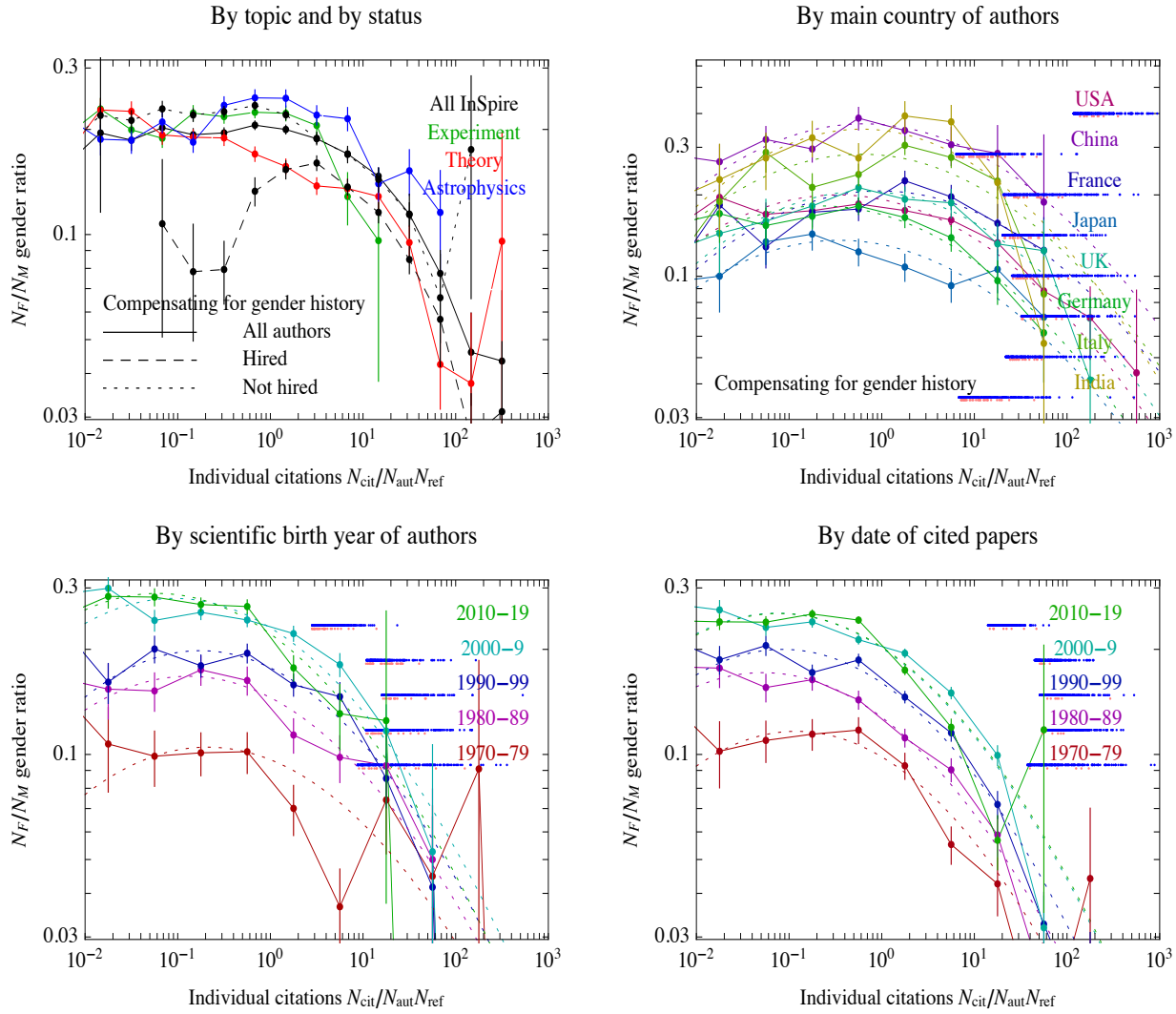

Figure S11: **Top-left:** *We compute the numbers N_F and N_M of female and male authors in fundamental physics who received a given amount $N_{\rm icit}$ of individual citations, and plot the ratio N_F/N_M (vertical axis) as function of* $\log N_{\rm icit}$ *(horizontal axis). The papers are here spitted according to their arXiv topic (all, experiment, theory, astrophysics) and the authors as all, hired, not hired (using 10yr-hires).* **Top-right***: same quantity splitting authors according to the most frequent country in their affiliations.* **Bottom-left***: same quantity splitting authors according to their scientific birth (publication year of the earliest paper), restricting to scientifically active authors up to the present decade.* **Bottom-right***: same quantity splitting papers according to their publication date. Data points with zero female or male authors are not shown; however in the latter 3 panels the points below the labels show the positions of the 300 top authors (M are blue, F are pink). The dotted curves show how eq. (8) with $p \approx 2$ and $R_\sigma \approx 2$ can fit the observed N_F/N_M.*

- The top-left panel of fig. S11 also shows that the trend is independently present among hired (dashed) and non-hired (dotted) authors. As expected, this distinction only affects the lower side of the $\log N_{\rm icit}$ distribution. We here used 10yr-hires. Furthermore the trend is independently present restricting to authors hired (or not hired) by top institutions, defined in various ways as discussed earlier.

- The fraction of M and F authors is approximatively the same within the countries that mostly contributed to fundamental physics: see the right panel of fig. 1. Thereby we do not expect that national differences are a confounder. Indeed, the top-right panel of fig. S11 shows that the trend is independently present dividing authors in non-overlapping samples according to the most frequent country in their affiliations (at least for large countries with enough statistics).

The plots above have been corrected for the generation effect as described in section 3.4; the plots below are restricted to time periods so that such correction is not necessary:

- The bottom-left panel of fig. S11 shows that the trend is independently present dividing authors in non-overlapping samples according to their scientific birth date (year of their first paper). We restricted to authors still scientifically active in the present decade (this restriction makes a little difference). This plot means that the trend is not due to physicists preferentially citing famous older authors who built their scientific reputation in years around 1970 when progress was faster and when the female fraction was lower. On the contrary, the trend is present also restricting to recent authors.

- The bottom-right panel of fig. S11 shows that the trend is independently present dividing papers in non-overlapping samples according to their publication date, since 1970 (when the InSpire data-base started being complete) up to now (recent papers are still accumulating citations).

S2 Statistical significance of the difference in variances

Statistical significance is a necessary condition for scientific significance but it is not a sufficient condition. A potential result is scientifically significant when the effect is larger enough than statistical uncertainties and than all other systematic uncertainties. Systematic uncertainties (confounding variables in the present study) are discussed in the main text.

We here address the statistical significance, following Appendix B of [Strumia et al.(2006)]. In order to quantify the significance of the gender gaps in productivity, and of their possible joint interpretation as a difference in upper variances, we consider the likelihood $\mathcal{L}$ that a statistical fluctuation of the background results into the observed apparent effect. When a model that interprets the effect is available, a sensitive statistics is the likelihood ratio between the null hypothesis and the model, quantified as a $\chi^2 = -2\ln\mathcal{L}$ difference. The $\Delta\chi^2$ can then be converted through statistical inference into a statistical significance (relative probability between the null hypothesis and the model, given the data) as well as into confidence regions for the model parameters. This is simply done in the Gaussian limit. Beyond this limit, statistical inference becomes a subtle and debatable subject that gives no objective univocal result when trying to quantify probabilities through frequentist or Bayesian techniques. The $\Delta\chi^2$ values we find are well above arbitrary thresholds for 'statistical significance' commonly considered in scientific disciplines, so that we don't need to address subtleties related to statistical inference.

We here quantify the statistical significance of the difference between the M and F binned distributions in fig. 13 or 14 through

$$\chi^2 = -2\sum_i \ln \mathcal{L}(\mu_{Mi}, N_{Mi}) P(\mu_{Fi}, N_{Fi}), \qquad \mu_{Gi} = N_G^{\rm tot} \frac{N_{Mi} + N_{F_i}}{N_M^{\rm tot} + N_F^{\rm tot}} \tag{7}$$

where $\mathcal{L}(\mu, N) = \mu^N e^{-\mu}/N!$ is the Poissonian likelihood of observing N events when μ are expected. We use Poissonian probabilities because some bins have zero or few authors. See [Bohm et al.(2014)] for Poissonian probabilities of weighted events; in the Gaussian limit the variance of $\sum_i w_i$ is $\sum_i w_i^2$.

We compute the χ^2 as a function of two fit parameters that artificially shift the mean and rescale the upper width of one of the two bells. We sum over bins i in the middle and upper parts of the bells, ignoring the lower tails. The middle part of the bells is included to reduce the uncertainties on the means, allowing to better see the difference in variances. We marginalise over the shift in means doing statistical inference in the common Gaussian limit of Bayesian and frequentistic techniques: in this limit marginalisation over parameters reduces to minimisation of the χ^2, and the final $\Delta\chi^2$ is approximatively distributed as a χ^2 with one degree of freedom. We find that the χ^2 decreases by a statistically significant amount ($\Delta\chi^2 \approx 50 - 100$ depending on how much of the middle region is included) when the width of the M distribution is artificially reduced by $10 - 15\%$.

We can also quantify the statistical significance through the approximated analysis that relies on eq. (5) and eq. (6) by fitting the N_F/N_M distributions plotted in fig. 15 to the following analytic form

$$R = R_N \exp\left(-\frac{R_\sigma}{2} \ln^p \frac{N_{\rm icit}}{N^0_{\rm icit}} \right) \tag{8}$$

where ln is the natural logarithm and R_N, R_σ, p, $N^0_{\rm icit}$ are free parameters. We then fit the observed N_{Fi}, N_{Mi} to the theoretical $R = N_F/N_M$ by defining a $\chi^2 = -2\sum_i \ln \mathcal{L}(R_i, N_{Fi}, N_{Mi})$ statistics, where the sum runs over the bins i, excluding the lower tails of the the $N_{\rm icit}$ distributions. Furthermore

$$\mathcal{L}(R, N_F, N_M) \propto R^{2N_F}(1+R)^{-2(N_F+N_M)} \tag{9}$$

(written up to an R-independent multiplicative factor) is the maximal likelihood of observing N_G authors with gender G assuming a given expected gender ratio R. We find that the fit parameter R_N depends on the overall number of M and F authors in the given sample; the fit parameter $N^0_{\rm icit}$ depends on the overall amount of citing literature relevant for the given sample. Different sub-samples are fitted reasonably well by common values of the extra fit parameters: $p \approx 2$ and $R_\sigma \approx 2 - 3$. Compensating for gender history the preference for $p = 2$ is again $\chi^2_{p=1} - \chi^2_{p=2} \approx 50 - 100$, depending how much of the middle region is included.