

# Do women get fewer citations than men?

*Sabine Hossenfelder*

In collaboration with Tobias Mistele and Tom Price.

Views expressed in this talk do not reflect or represent FIAS, FQXi, or the DFG.

# The brief answer...

... is “yes.”

But hear me out before jumping to conclusions.

# How did I ever get into this?

I am an ex-particle-physicist with an interest in bibliometric analysis.

I have recently done a few studies using data from the arXiv open access server. For example, we trained a neural net to predict citation counts. That was fun. We want to offer this prediction online in the soon future (more about this later).

Then this happened


# Then this happened



## Cern scientist: 'Physics built by men - not by invitation'

By Pallab Ghosh  
Science correspondent, BBC News

🕒 1 October 2018

     Share

# Then this happened

**BBC** Sign in News Sport Weather Shop Reel Travel

**NEWS**

Home Video World UK Business Tech Science Stories Entertainment & Arts

World Africa Asia Australia Europe Latin America Middle East US & Canada

## Cern scientist: 'Physics built by men - not by invitation'

By Pallab Ghosh  
Science correspondent, BBC News

🕒 1 October 2018

f WhatsApp Twitter Email Share

IOP Publishing View our other sites

physicsworld 🔍 Magazine | Latest ▼ | People ▼ | I



CULTURE, HISTORY AND SOCIETY | NEWS

### CERN describes physicist's gender talk as 'highly offensive'

01 Oct 2018 Hamish Johnston



Collision course: CERN has condemned a talk given by an Italian particle physicist. (Courtesy: Maximilien Brice/CERN)

Alessandro Strumia claimed "physics invented and built by men"

# Then this happened



The screenshot shows the top section of The Guardian's website. On the left, there's a red 'BBC NEWS' logo with 'Home' and 'Video' links. Below it are 'World' and 'Africa' links. The main header is dark blue with 'Support The Guardian' in yellow, followed by 'Contribute' and 'Subscribe' buttons. A 'Sign in' link is on the right. The Guardian logo is prominently displayed. Below the header is a navigation bar with 'News', 'Opinion', 'Sport', 'Culture', and 'Lifestyle' in white. A yellow menu icon is on the right. Underneath is a secondary navigation bar with 'World', 'Europe', 'US', 'Americas', 'Asia', 'Australia', 'Middle East', 'Africa', 'Inequality', 'Cities', and 'More'. The main content area features a red 'Cern' tag, a large headline 'Physics was built by men': Cern suspends scientist over remarks', and a sub-headline 'Italian professor's presentation deemed 'unacceptable' by Geneva research centre'. The author 'By Pallab Ghosh, Science correspondent' and the date '1 October 2018' are on the left. A partial quote from Alessandro Strumia is at the bottom right.

**BBC NEWS**  
Home | Video  
World | Africa

**Support The Guardian**  
Contribute → Subscribe →  
Sign in

**The Guardian**

News | Opinion | Sport | Culture | Lifestyle

World ► Europe US Americas Asia Australia Middle East Africa Inequality Cities More

**Cern**

**Cern sci by invita**  
By Pallab Ghosh  
Science correspondent  
1 October 2018

**'Physics was built by men': Cern suspends scientist over remarks**

**Italian professor's presentation deemed 'unacceptable' by Geneva research centre**

Alessandro Strumia claimed "physics invented and built by men"

# Then this happened

**BBC** **NEWS**

Home | Video

World | Africa

**physicsworld** IOP Publishing View our other sites Magazine | Latest | People


Thousands of physicists sign letter condemning 'disgraceful' Alessandro Strumia gender talk

08 Oct 2018 Michael Banks

**Cern sci by invita**

By Pallab Ghosh  
Science correspondent

1 October 2018



Delegates at the first workshop on high-energy theory and gender held at CERN last month. (Courtesy: CERN)

Over 3000 physicists denounce an invited talk given at last month's inaugural CERN workshop on high-energy theory and gender

Sign in

# The Guardian

Culture | Lifestyle

lia Middle East Africa Inequality Cities More

## men': Cern suspends

### l 'unacceptable' by Geneva

Alessandro Strumia claimed "physics invented and built by men"




# Then this happened

**BBC**  
**NEWS**  
Home | Video  
World | Africa

**physicsworld**  
Magazine | Latest | People | I

**physicist**  
Thousands of physicists sign letter condemning 'disgraceful' Alessandro Strumia gender talk  
08 Oct 2018 Michael Banks



Delegates at the first workshop on high-energy theory and gender held at CERN last month. (Courtesy: CERN)

**Cern sci**  
**by invita**  
By Pallab Ghosh  
Science correspondent  
1 October 2018


Over 3000 physicists denounce an invited talk given at last month's inaugural CERN workshop on high-energy theory and gender

**THE INDEPENDENT**  
JOIN US | SUBSCRIBE / REGISTER | LOGIN

**PHYSICIST WHO CLAIMED 'PHYSICS WAS INVENTED BY MEN' SUSPENDED BY CERN WITH IMMEDIATE EFFECT**

mod up in front of a room of early career women and delivered a manifesto on why women didn't belong in physics'

**ia Barr** | @fabsab5  
Tuesday 2 October 2018 09:50 | 5 comments

  
Click to follow  
Indy/Life

physicist who came under fire for claiming that physics was “invented and built by men” while giving a talk at a European Organisation for Nuclear Research (Cern) workshop in Geneva has been ended from the organisation with immediate effect.

essor Alessandro Strumia, a senior researcher at the National Institute of Chemical Physics and physics, had allegedly asked Cern if he could deliver a talk on his most recent bibliometrics r.

Alessandro Strumia claimed “physics invented and built by men”

# The Strumia talk

I did not attend the talk. I have only seen slides.

The slides are unprofessional and unnecessarily offensive.

Still, his point that papers with female authors are systematically less cited in high energy physics warrants attention.

And I had my hands on the data anyway.

# Claim: Women get cited less than men

x-axis shows time since first paper on Inspire (people to the right are older)

y-axis shows cumulative citations normalized to number of authors

**That's a HUGE difference!**

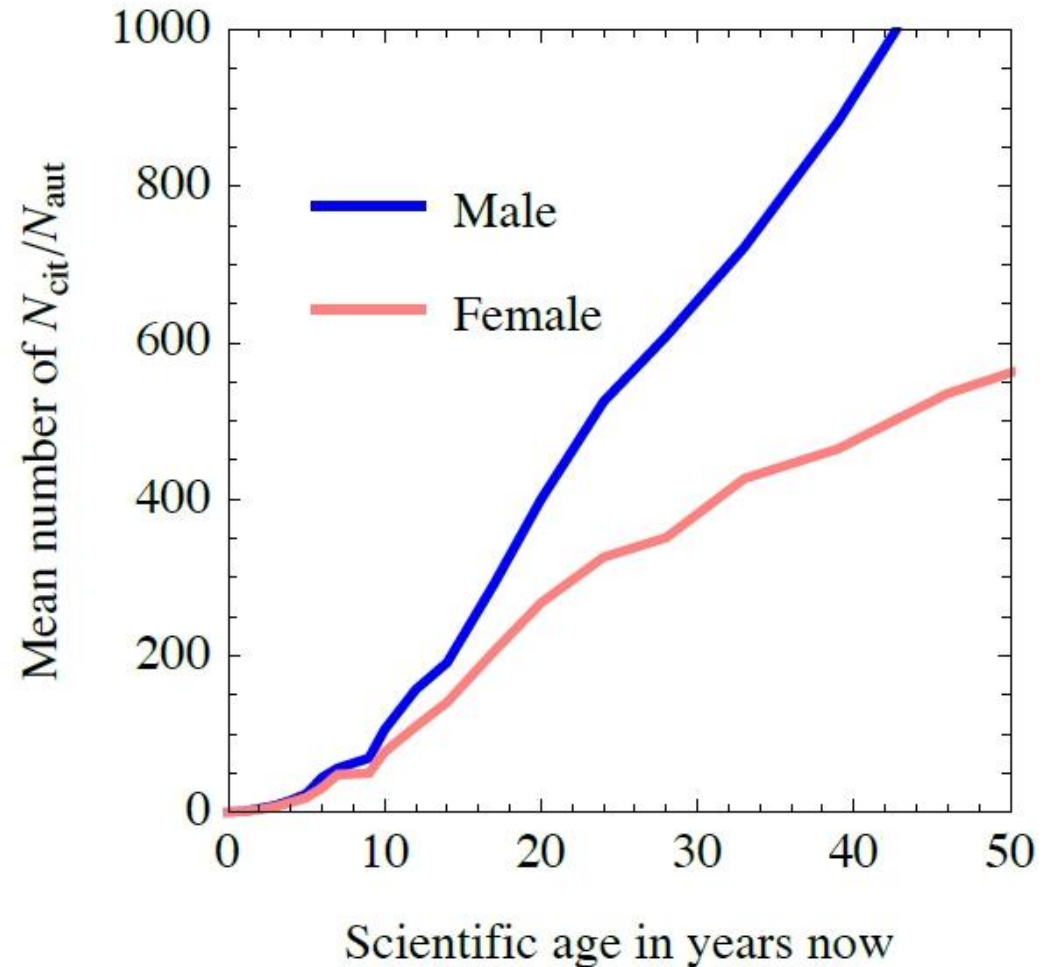


Figure: Alessandro Strumia

# Can we reproduce this finding?

We previously used a different data set.

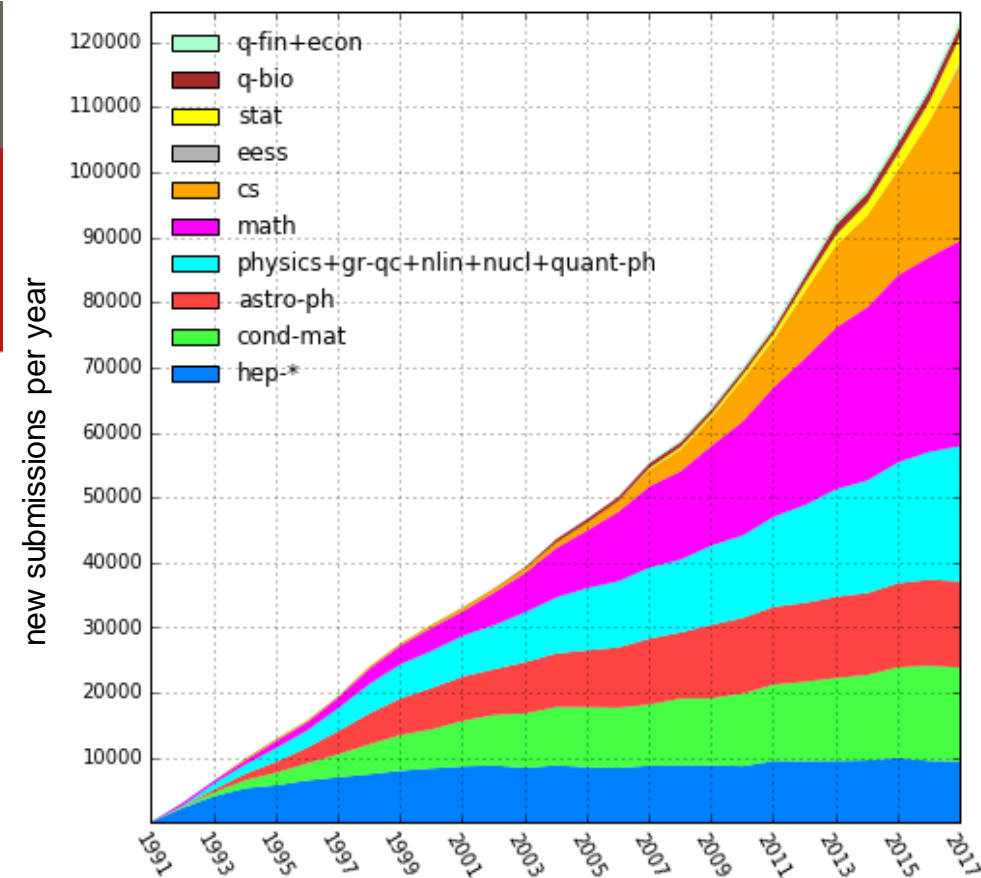
Rather than using the Inspire data base, we used the arXiv data plus citation data from Paperscape.

What are the main differences?



- Open access server
- In operation since 1994
- We use the “physics” set only.
- Dominated by hep, astro-ph, and cond-mat
- Papers up to April 2018

Citation data from Paperscape: In-arXiv references only. Coverage reaching more than 80% for the “early” fields, but less than 10% for “late” fields. Data up to December 2017.





Inspire also offers other services to the community, such as conference listings, job postings, and it collects information about institutions and authors.

- Formerly SPIRES.
- Digital library/content management system.
- About 1.1 mio papers dating back to the early 20<sup>th</sup> century
- Only particle physics and closely related disciplines

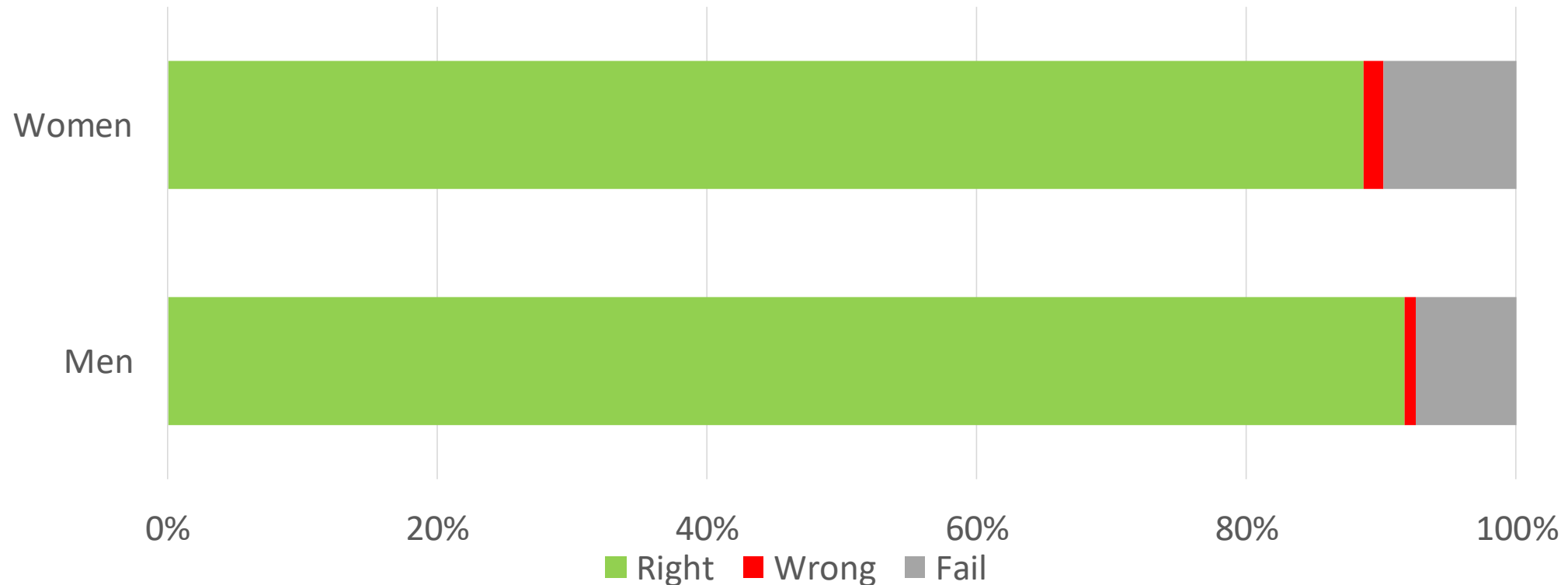
Since the number of publications increases exponentially and the arXiv is dominated by the fields that are covered by Inspire, most papers in either database have a corresponding entry in the other database. I.e, results from analyzing one of those databases, while not quantitatively identically, should be comparable with the other.

# Gender-identification

- We use a publicly available list of international first names (the same list as Strumia *et al*)
- We match names to author names using the algorithm described in our earlier paper, arXiv:1805.04647 [physics.soc-ph].
- Authors whose name are ambiguous (Simone, Andrea, Stacy, etc) are marked as not identified and are removed from the sample.
- We did at this point not attempt to further improve on this algorithm by using affiliations or email addresses.

# Gender-identification check

We tested our gender-id algorithm using lists of participants of the recent Boston Marathons (2015/2016/2017). Ca 80,000 international participants with (self-supplied) male/female identification.





# How good is our gender-identification?

Misidentification about 1% for men and 2% for women.

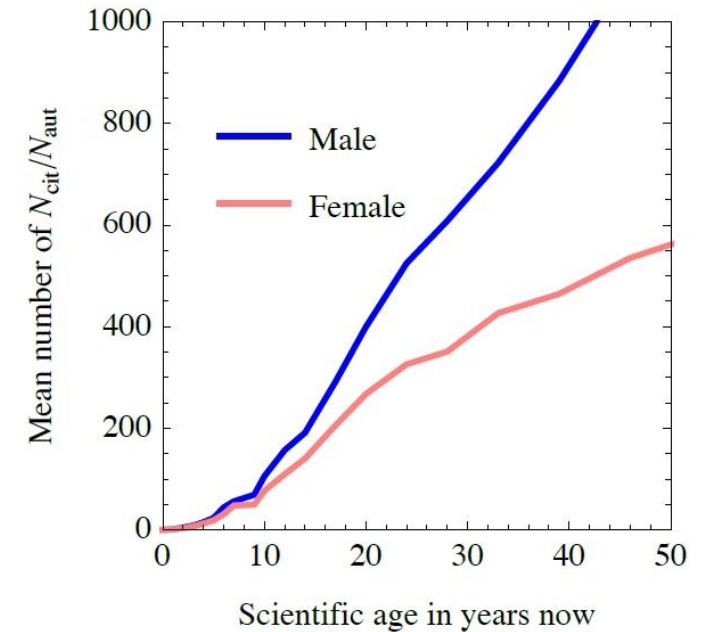
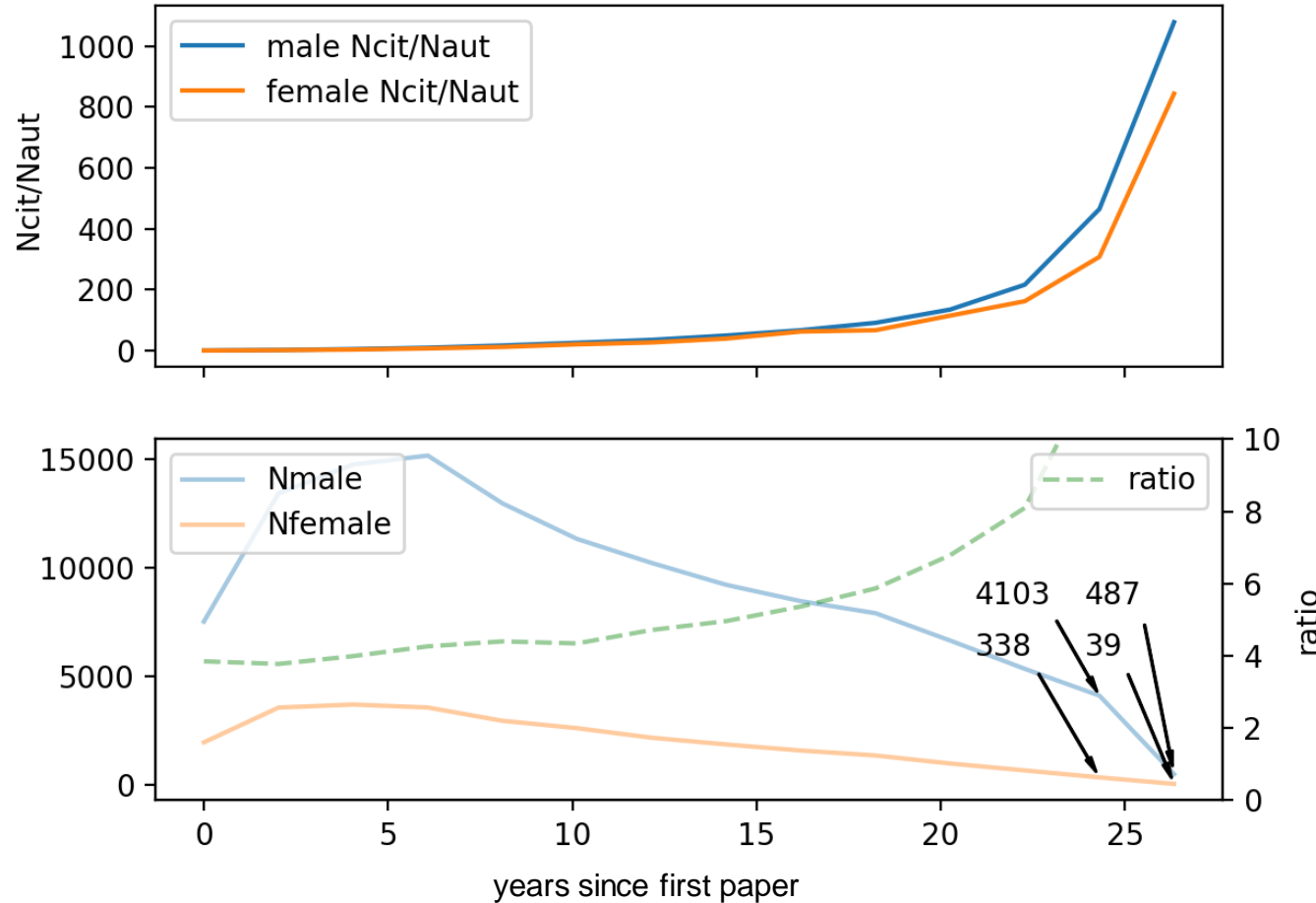
Gender-identification is somewhat more likely to fail for women than for men, meaning that in the following **total numbers slightly underestimate the true fraction of women.**

# Taking into account the leaky pipe

In the following, we remove “inactive” researchers, which are researchers who have not published a paper past 01-01-2015 or have fewer than 5 papers in total.

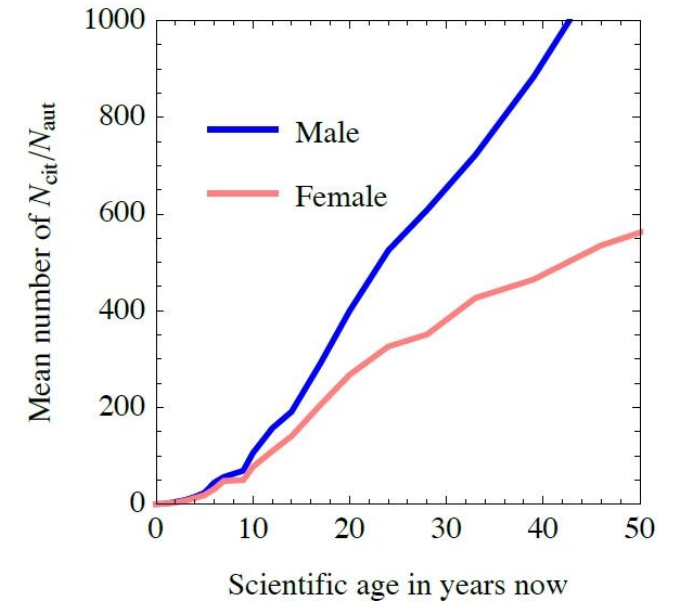
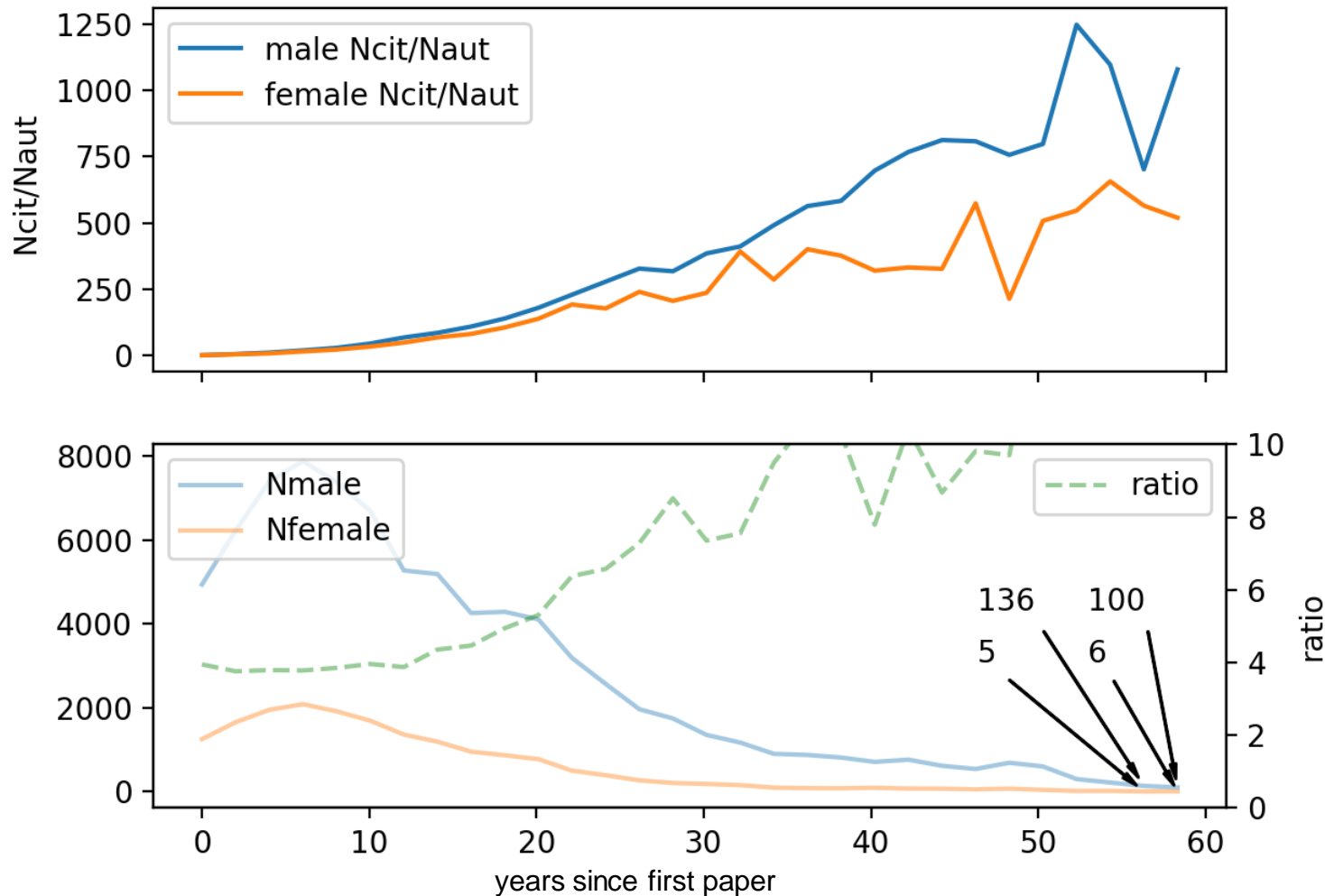
This is to avoid that the higher rate of dropouts among women pulls down the female average.

# 1<sup>st</sup> Try: Repeat analysis with ArXiv data



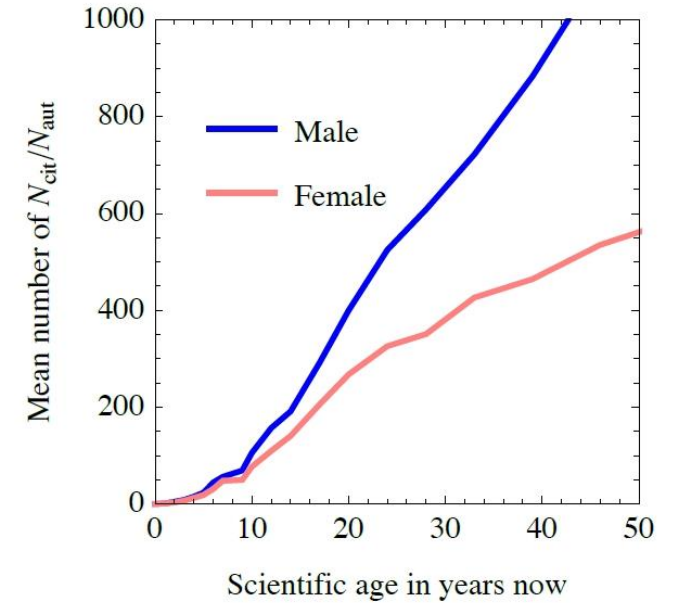
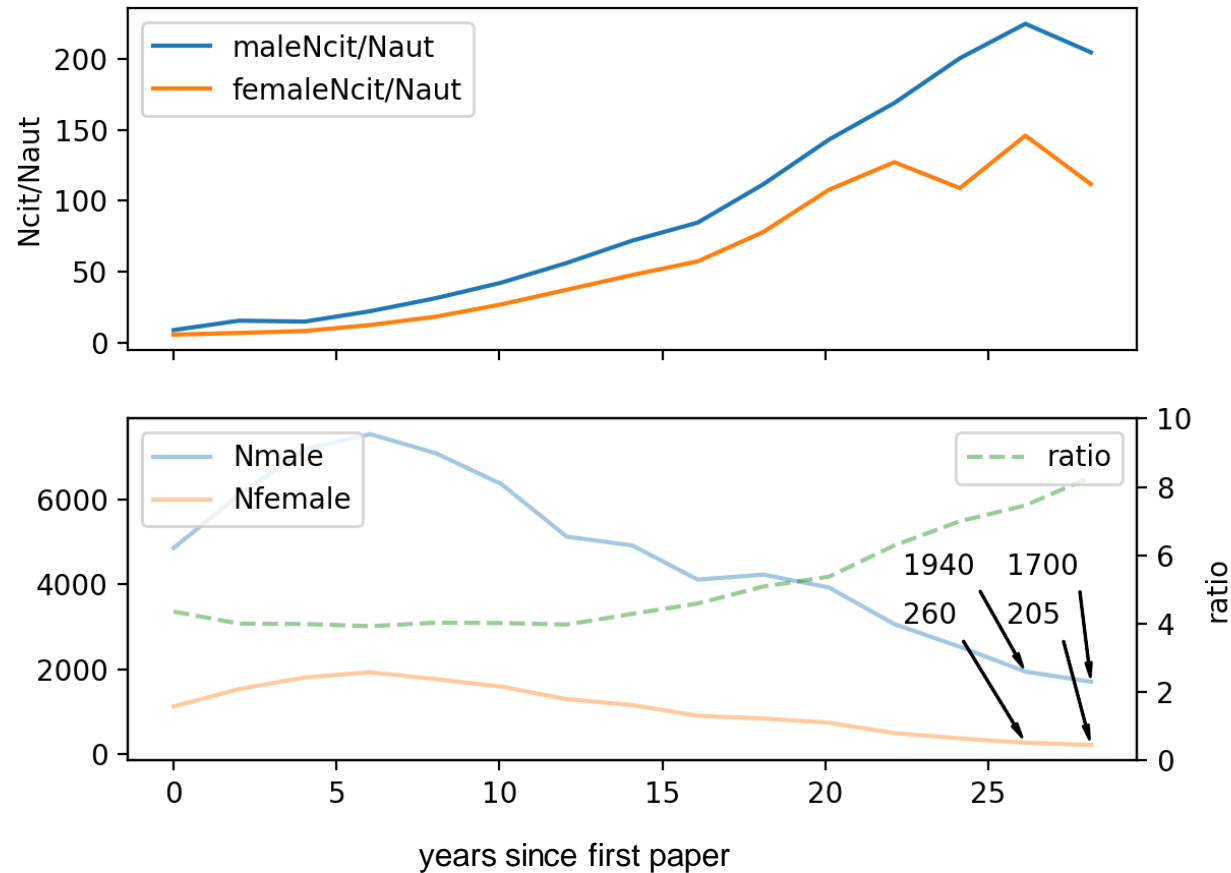
Note: The shape of the curve differs from the Inspire curve b/c the arXiv curve also integrates over the growth of the arXiv itself.

# 2<sup>nd</sup> Try: Repeat analysis with Inspire data



**Note: Data from papers older than 40 years very noisy due to small sample.**

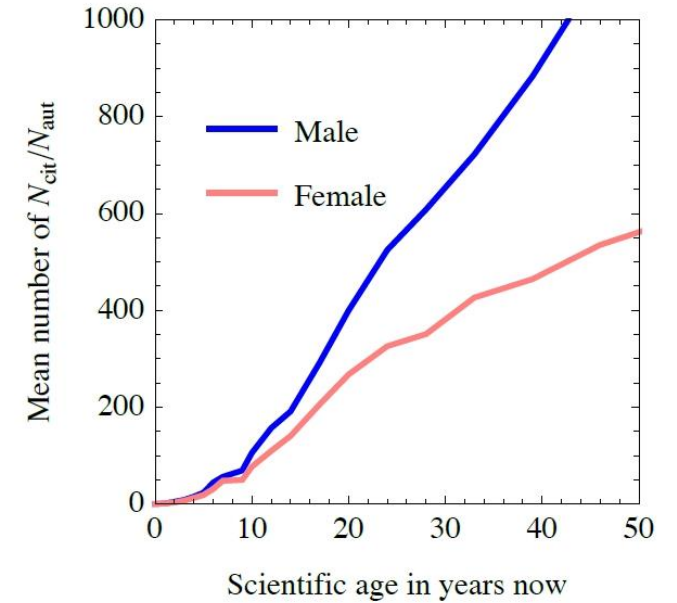
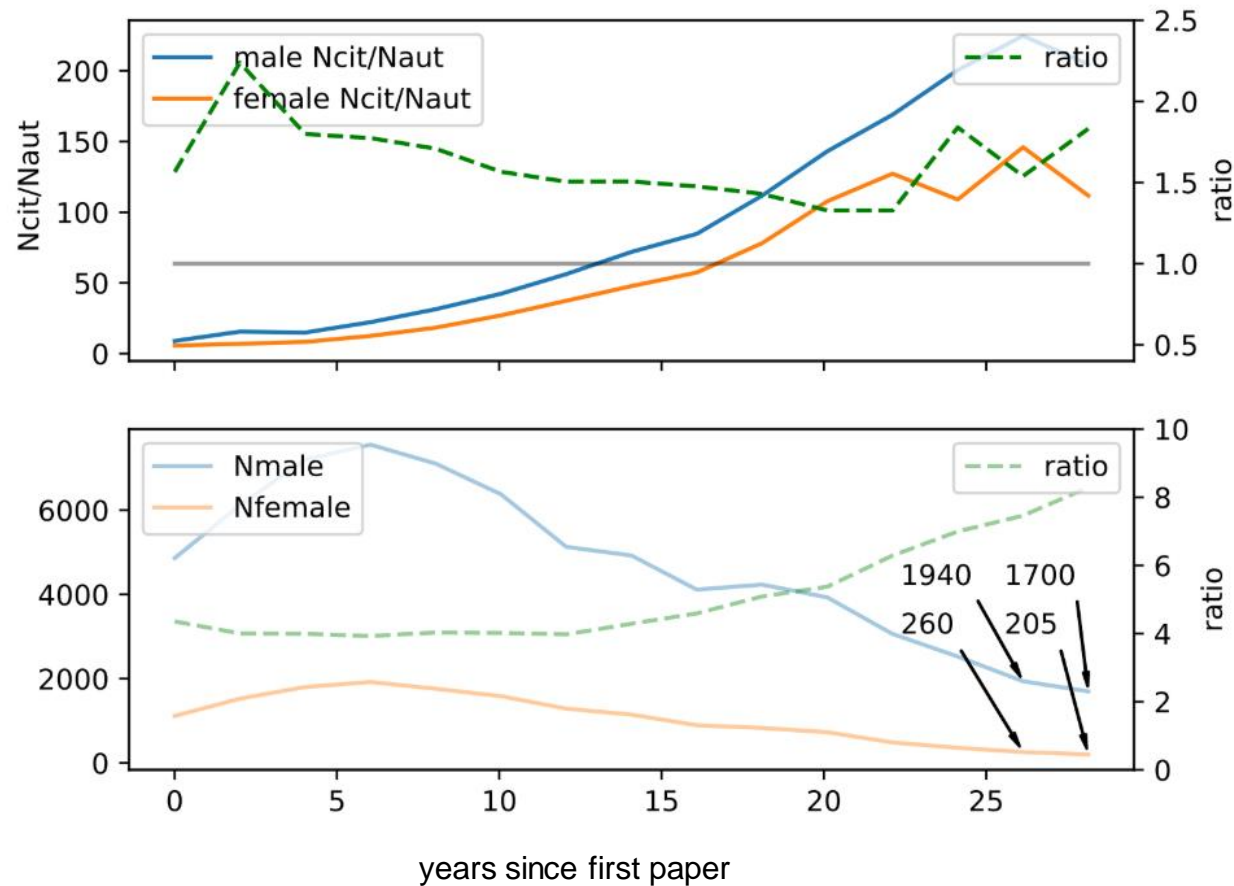
# 3<sup>rd</sup> Try: ArXiv data with Inspire-age



The results are now largely comparable:

The (current) male/female ratio is about 1.5 both for the arXiv and Inspire.

# 3<sup>rd</sup> Try: ArXiv data with Inspire-age



The results are now largely comparable:

The (current) male/female ratio is about 1.5 both for the arXiv and Inspire.

# Where did the huge difference go?

The gender-difference is most pronounced around the (current) scientific age of 50 (larger than 2), which the arXiv data does not contain. Even the biggest arXiv-difference is lower (about 1.5).

Using arXiv data only underestimates the “scientific age” for people whose first paper predates the arXiv. Correcting this *increases* the difference.

Reason: The gender-difference happens to be small in the years directly before the arXiv launched, and these are the most populated years. (Why was the difference small at the time? We have no idea.)

# So: The gender-difference exists and is robust

- A difference in the citations to male and female authors is clearly present both in the arXiv and in the Inspire data.
- The difference is significantly larger than the estimated error from the gender-identification.
- The difference remains up to 2017. It is not an “old-age” effect.



# Where does the difference come from?

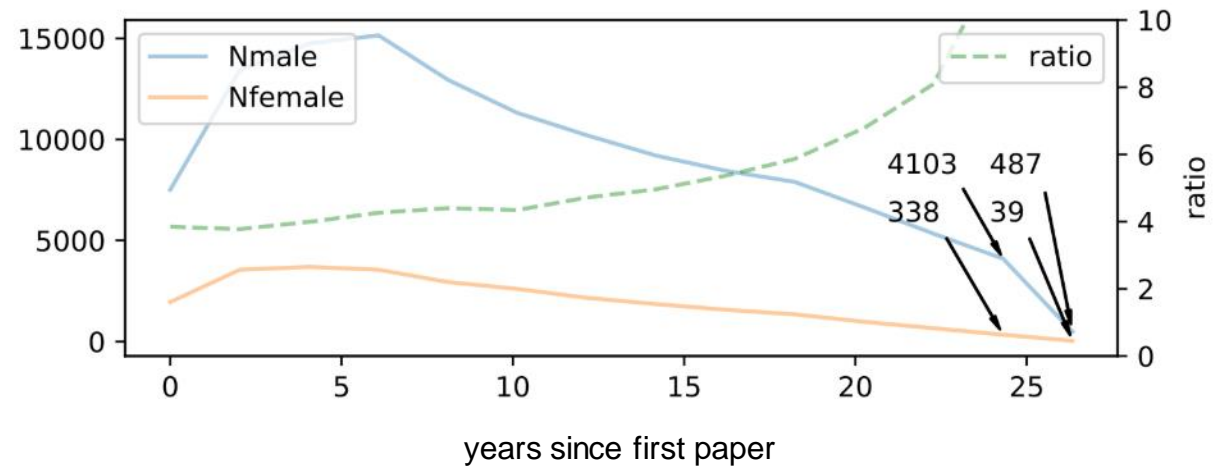
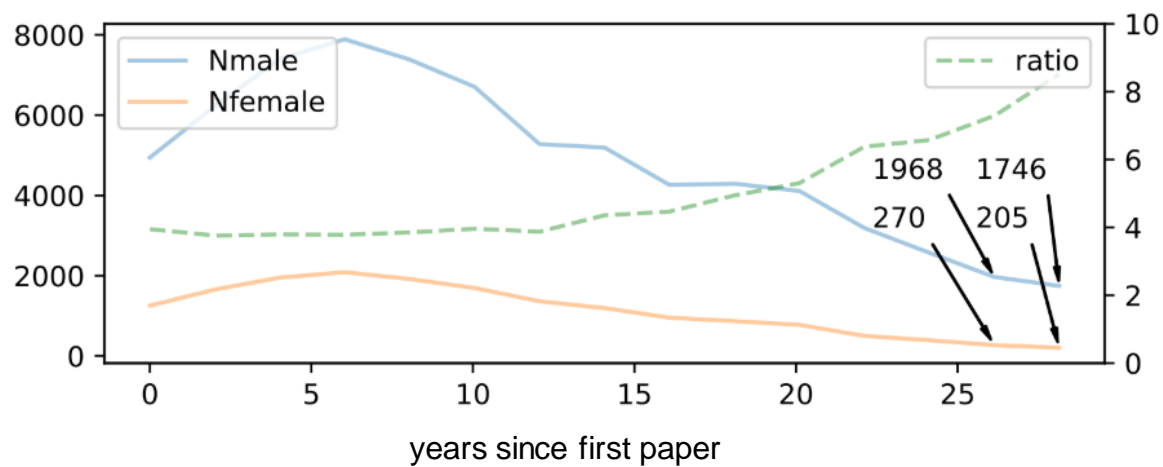
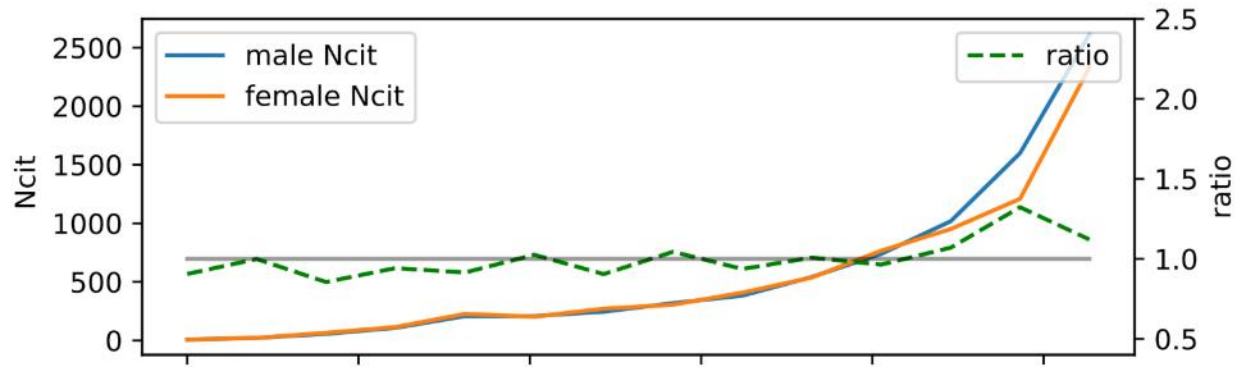
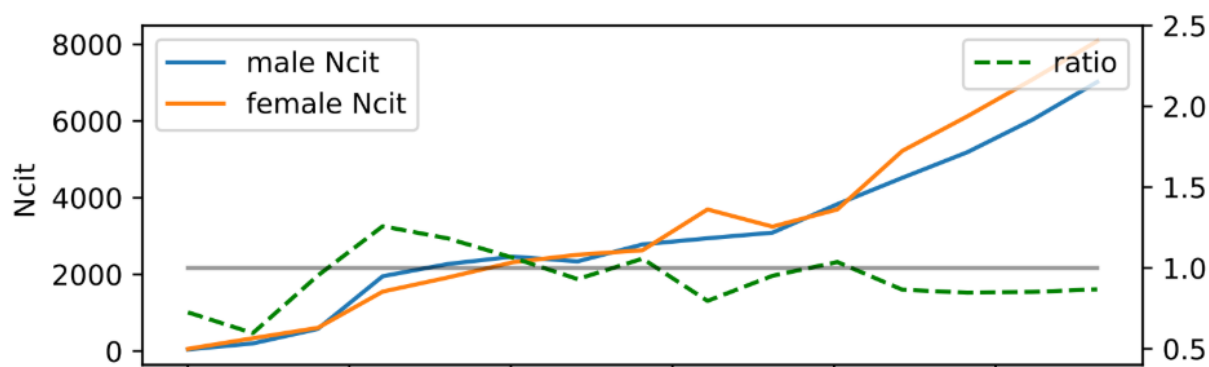
Normalizing the number of citations to the number of authors is not a commonly used measure.

**Advantage:** For small collaborations, it is a better first approximation to the actual work that each author did (and hence credit they should get).

**Disadvantage:** Makes it difficult to tell the origin of the gender-difference.

Citation analyses for members of large collaborations is meaningless because in this case individual contributions cannot be extracted from authorship. For this reason, large collaborations are normally excluded from citation analyses. Normalizing the number of citations by the number of authors makes such papers irrelevant, so that further explicitly removing them does not change much.

# The total number of citations is comparable



Left: Inspire, Right: arXiv

# Do women collaborate more frequently?

Normalizing the number of citations ( $N_{\text{cit}}$ ) to the number of authors ( $N_{\text{aut}}$ ) has the effect of making single-authored papers most relevant.

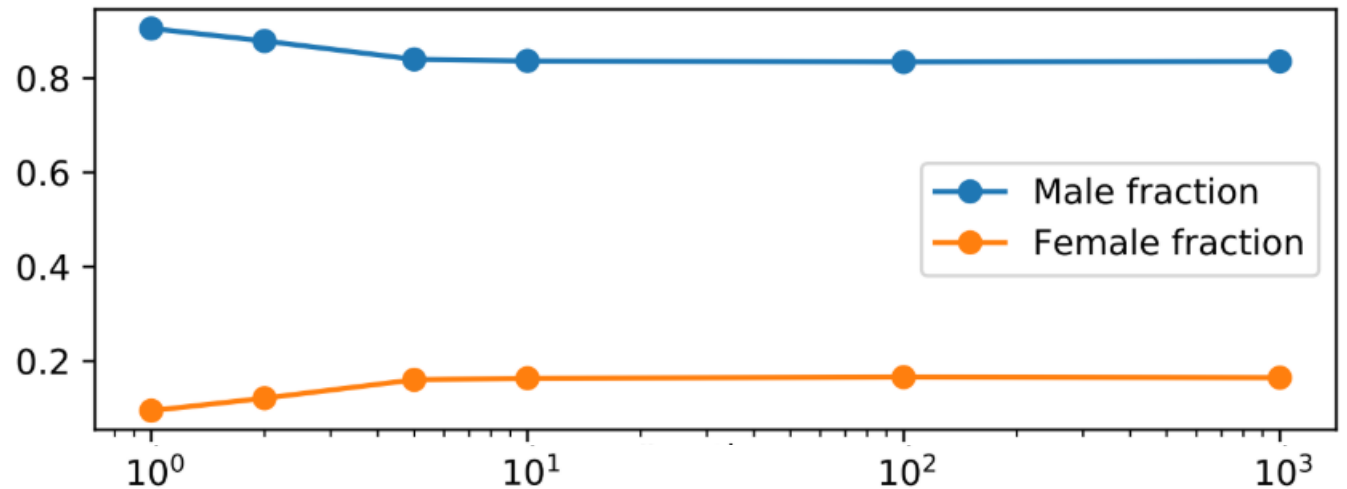
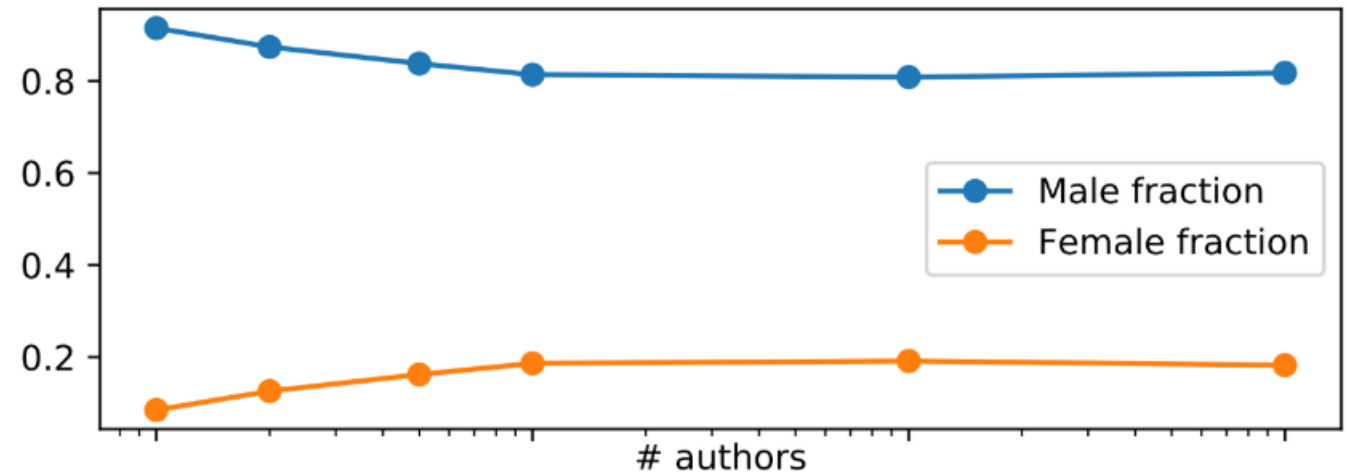
The total number of citations to all single-authored papers however also depends on the number of single authored papers (duh).

This brings up the question: Does the difference come from the frequency by which women and men publish single authored papers?

# Who collaborates how much?

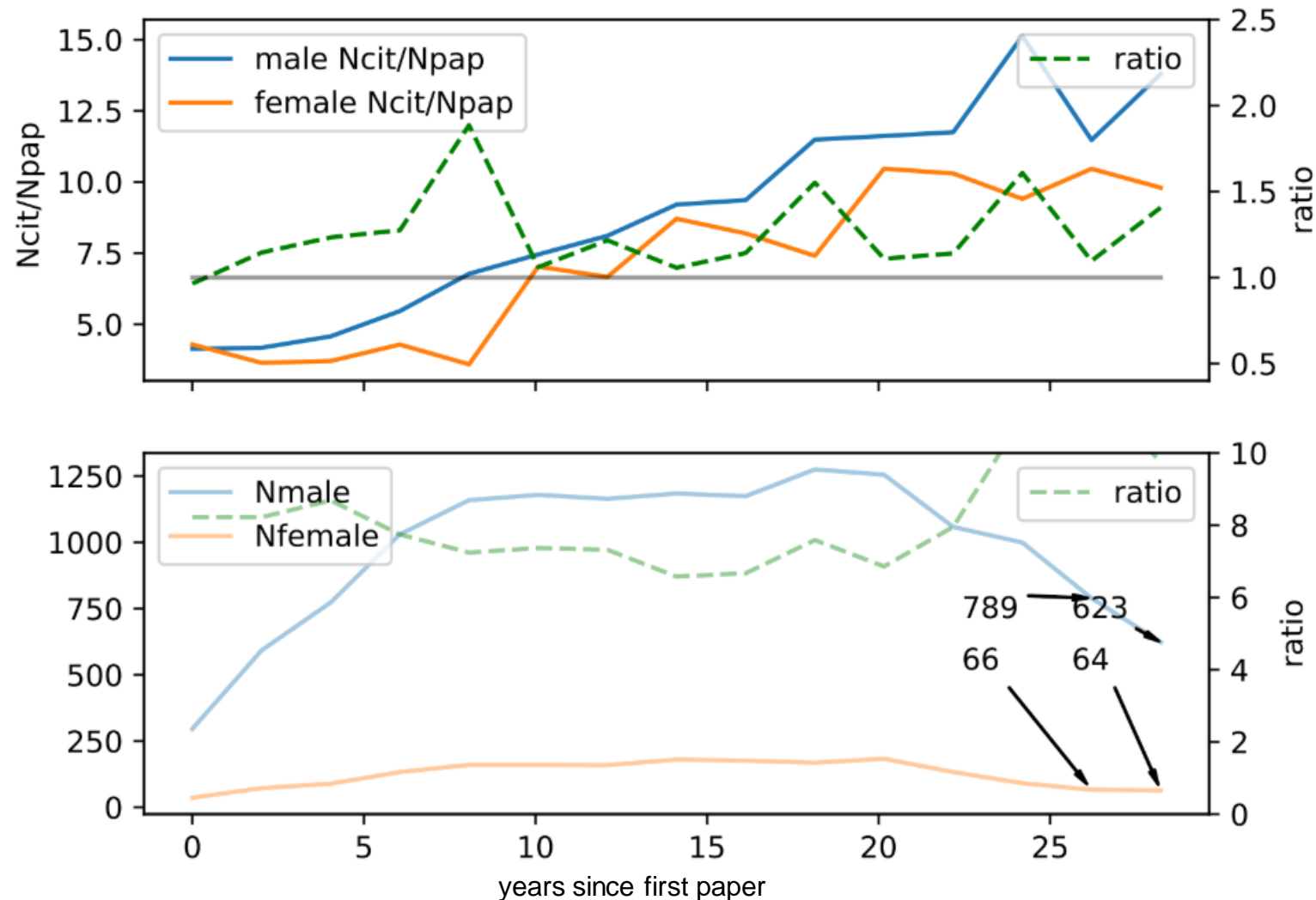
Women publish (relatively) fewer single-authored papers than men and work more in small collaborations.

This explains part of the difference in  $N_{\text{cit}}/N_{\text{aut}}$ .



Top: Inspire, Bottom: ArXiv

# Citations to single-authored papers only



The number of citations to individual single-authored papers by women is also smaller than that to single-authored papers by men.

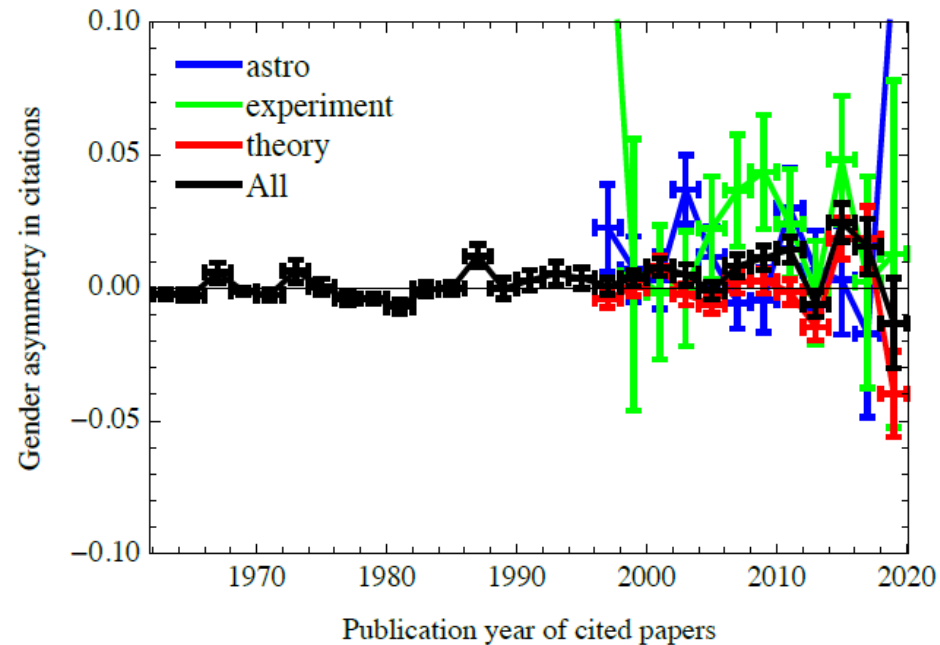
(Figure: arXiv data with Inspire age.)

# Where do the citations come from?

Count single-author that cites different single-author, define

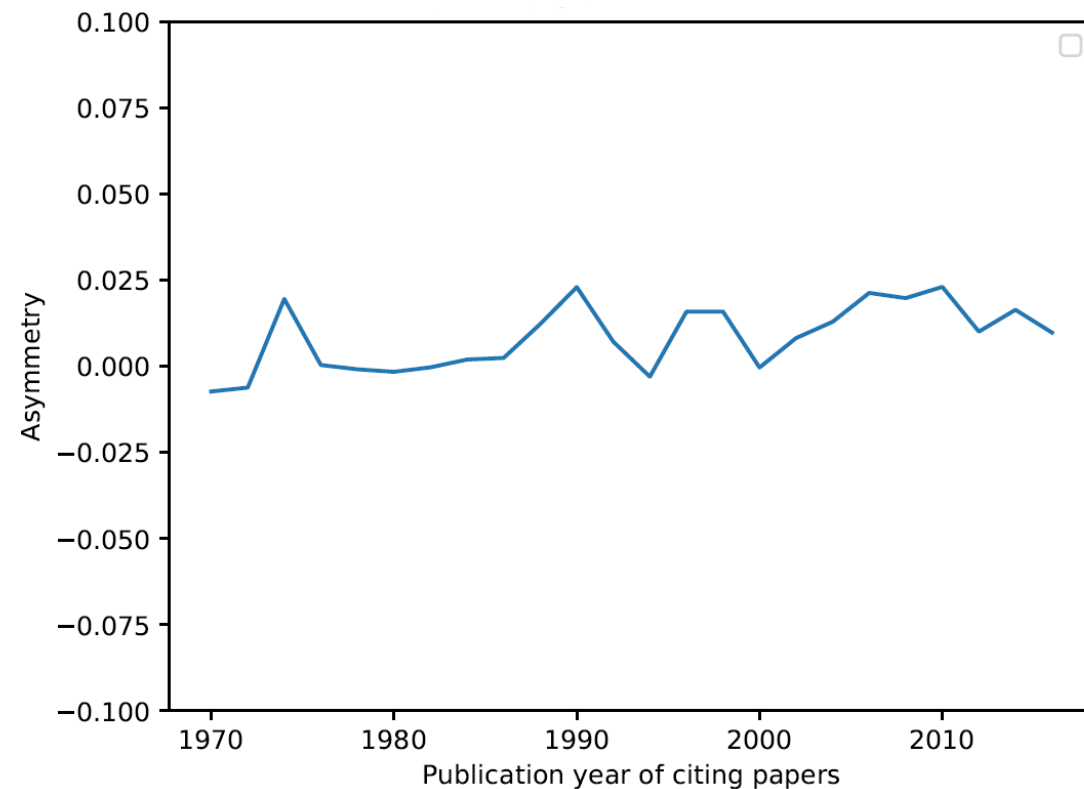
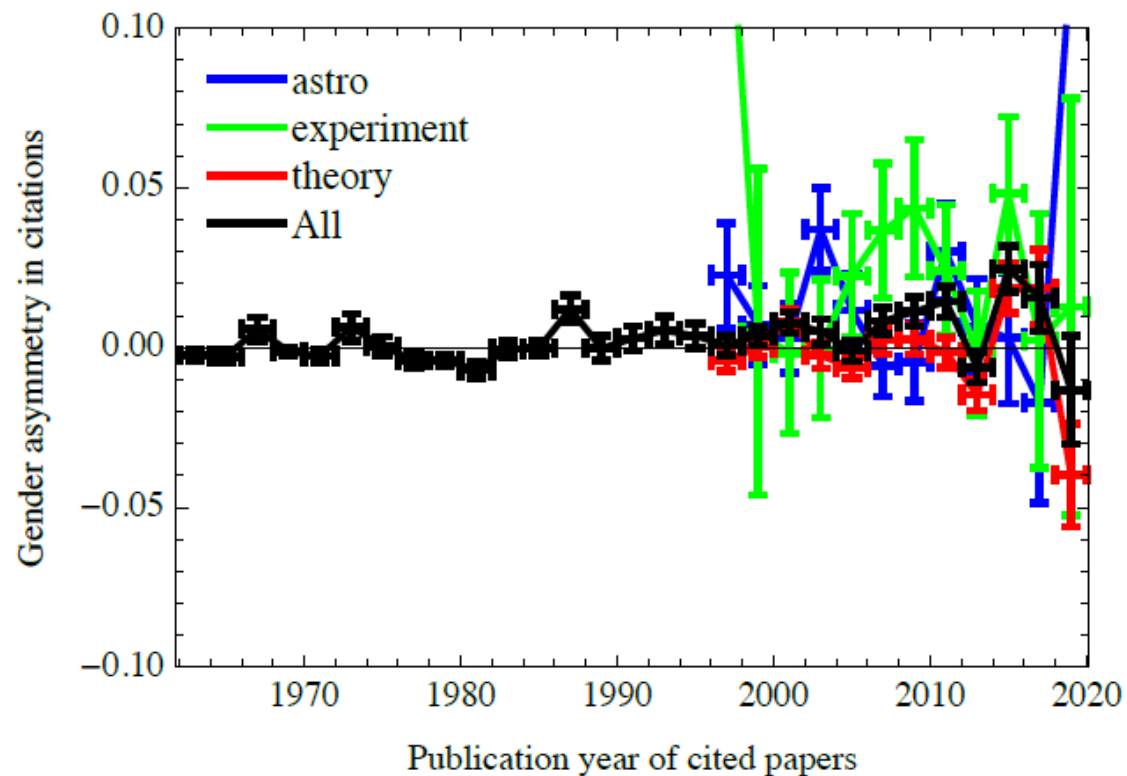
$$\text{Gender asymmetry} = A \equiv \frac{1}{N_{M \rightarrow}^{\text{cit}} N_{F \rightarrow}^{\text{cit}}} \det \begin{pmatrix} N_{M \rightarrow M}^{\text{cit}} & N_{M \rightarrow F}^{\text{cit}} \\ N_{F \rightarrow M}^{\text{cit}} & N_{F \rightarrow F}^{\text{cit}} \end{pmatrix} \quad -1 \leq A \leq 1$$

InSpire, after 2010	
category	$A$
hep-ex	$(-1.2 \pm 1.7)\%$
hep-ph	$(0.1 \pm 0.6)\%$
hep-th	$(-0.1 \pm 0.7)\%$ ,
astro-ph	$(0.6 \pm 1.1)\%$
hep-lat	$(0.2 \pm 2.2)\%$
nucl-ex	$(0.2 \pm 2.1)\%$
gr-qc	$(0.2 \pm 1.2)\%$



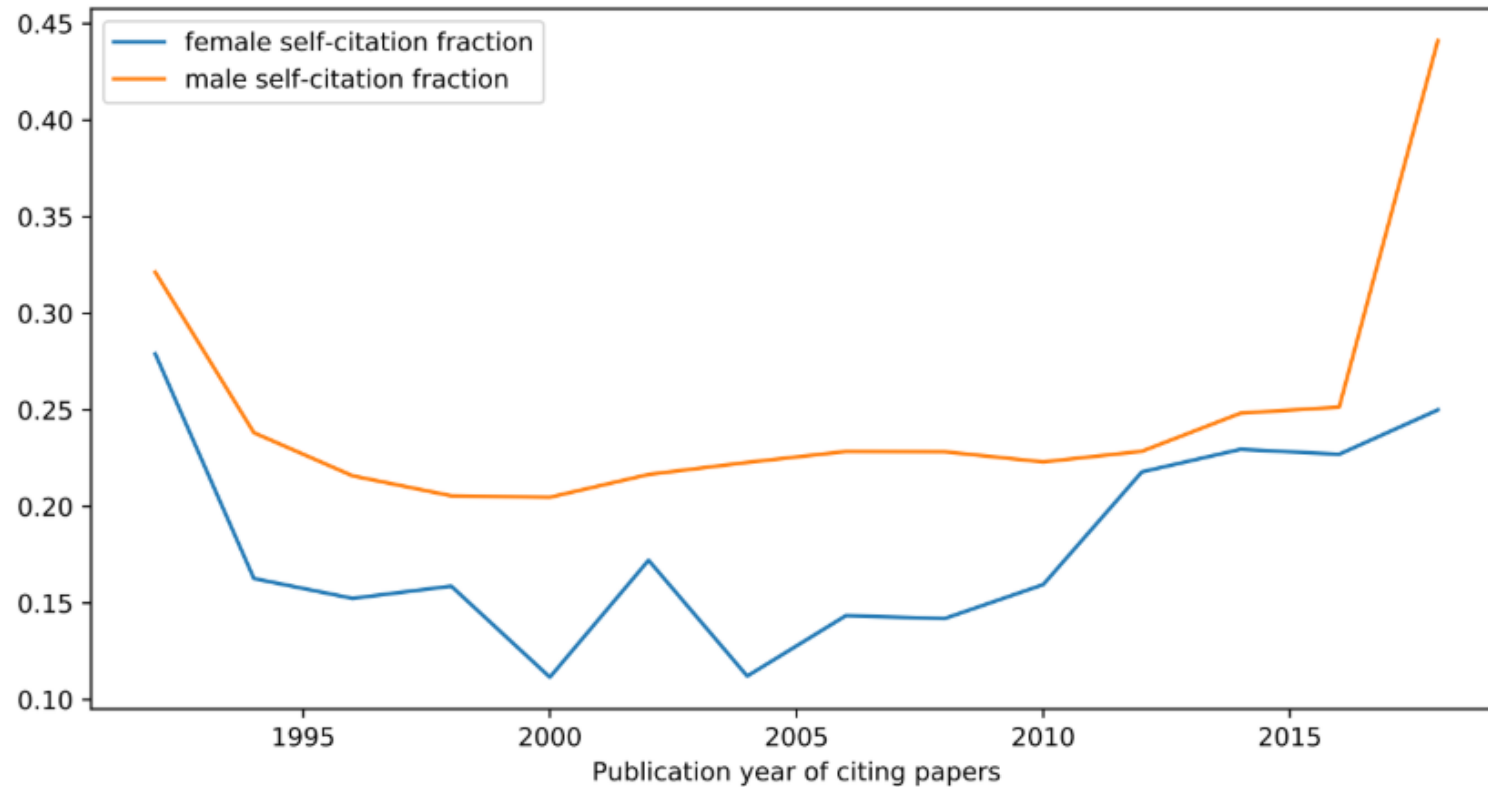
No gender preference in citations in any category at any time, down to % level.

# We can reproduce this\*



\* if we exclude self-citations

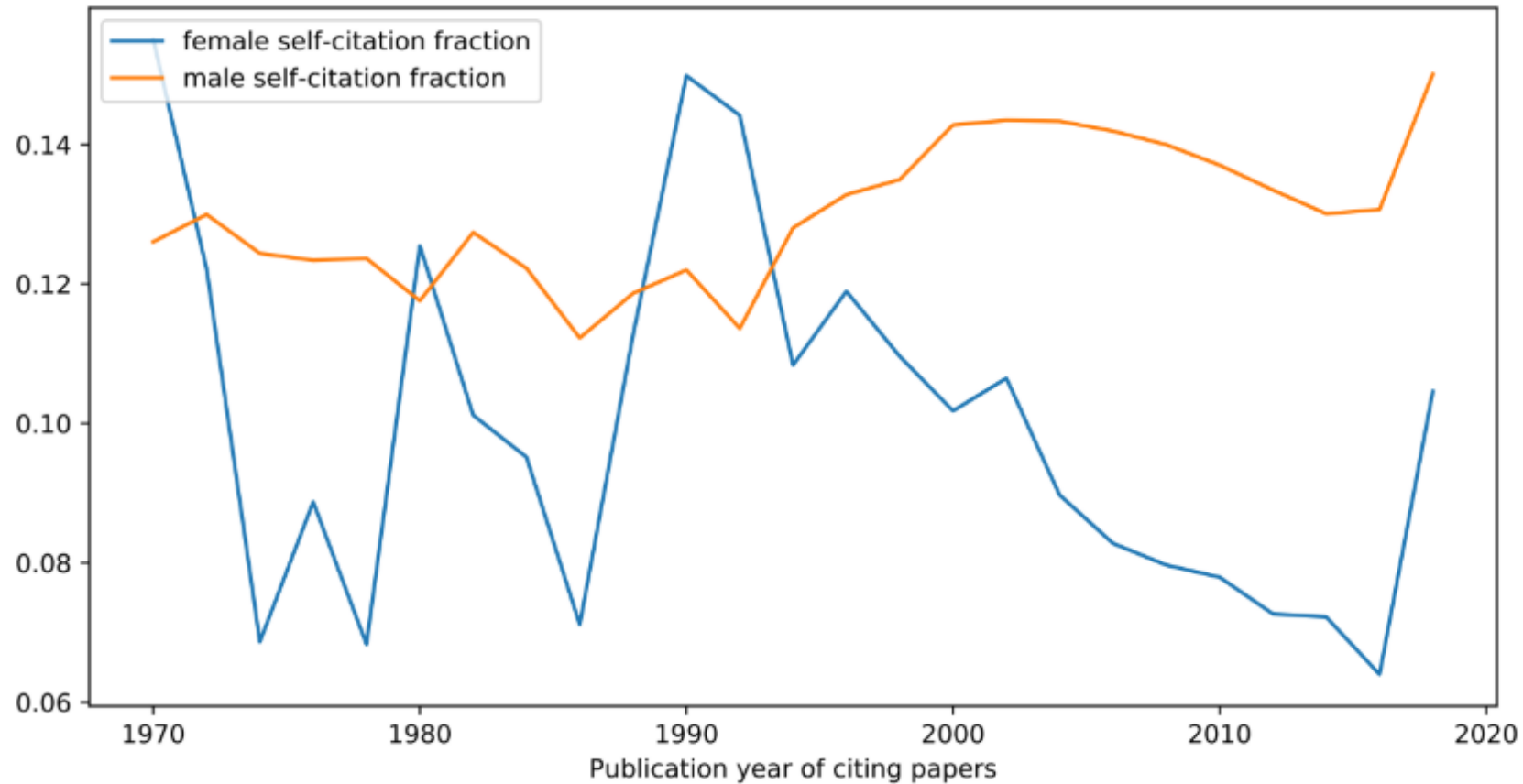
# Fun fact: Men cite themselves more often



arXiv data



# Fun fact: Men cite themselves more often



Inspire data

# What do the self-citations do?

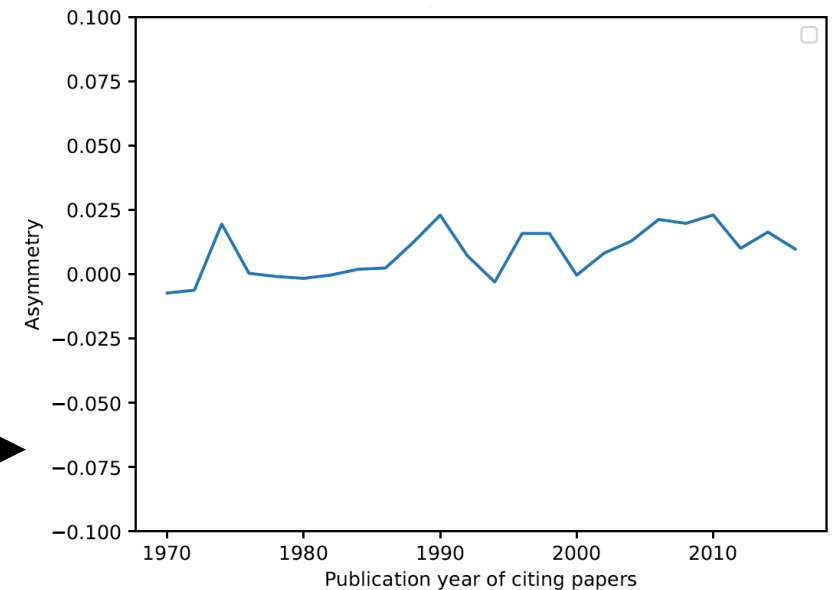
Including self-citations significantly increases the gender-asymmetry.

Self-citations however have a negligible effect on the  $N_{\text{cit}}/N_{\text{aut}}$  ratio.



← with self-citations

without self-citations →



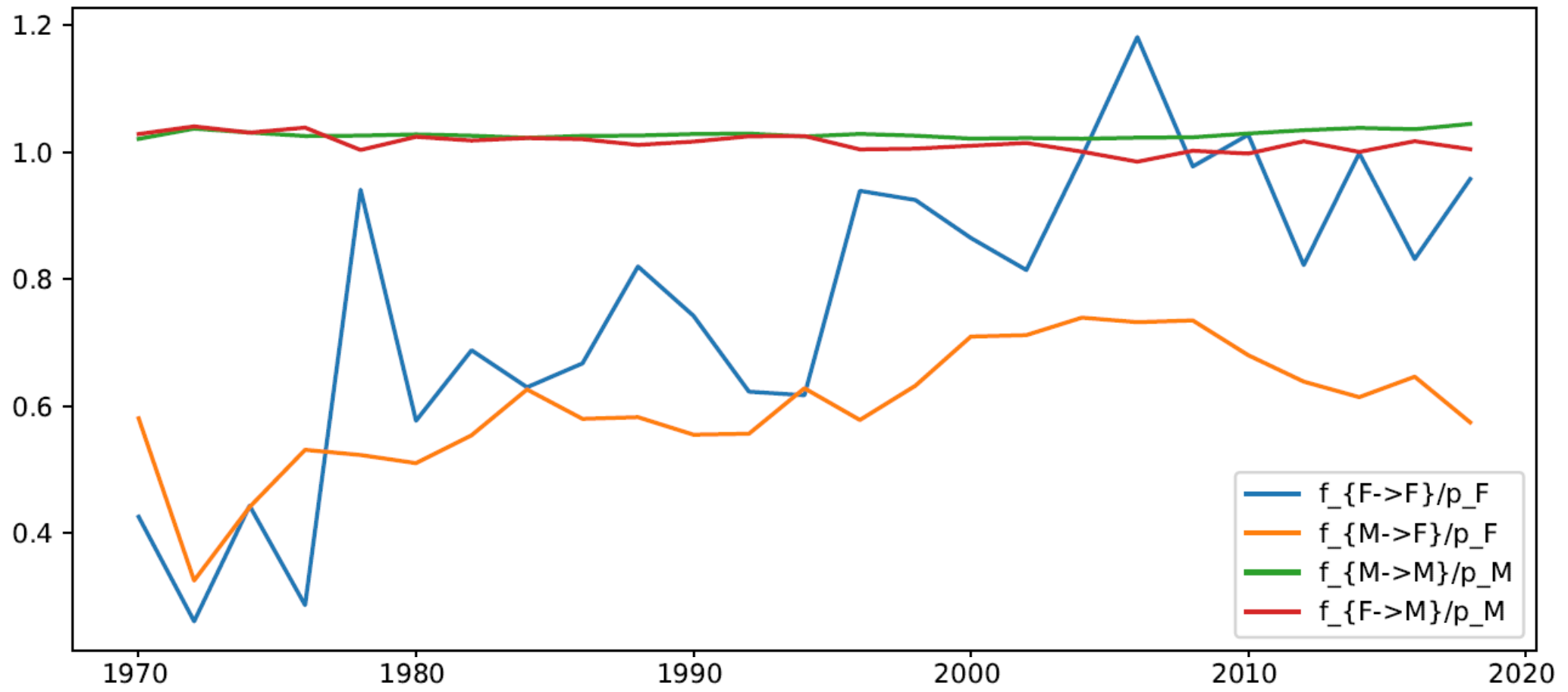
# But what does this asymmetry mean?

Unclear what it means that deviations of determinant from zero are below percent level. Is this small? Is this large?

To obtain a more meaningful comparison:

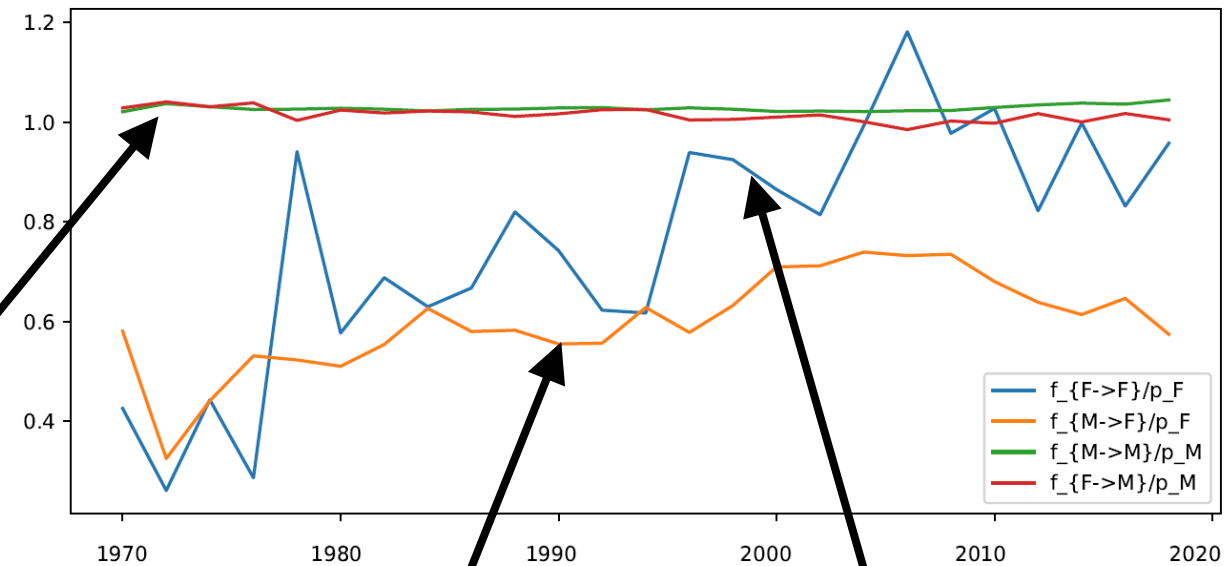
Look at fraction of citations given to men/women relative to “random pick”: Divide ratio of citations from M/F to M/F (notated  $f_{M \rightarrow M}$  etc) to fraction of M/F papers in the pool of all citable papers ( $p_M$  or  $p_F$ ).

# Who cites whom?



# What does this mean?

- Men and women both cite men roughly with the frequency that one would expect from a gender-mixture.
- Men cite women consistently less than a gender-mixture would suggest.
- Women used to cite women less, but in the past 20 years or so have taken on to citing women roughly at the “gender-mixed” level. (Note though that fluctuations are large, so the trend may be spurious.)



# This difference is not large but...

... it has a large effect on the citations to female papers because there are so many more men than women.

# To recapitulate:

The ratio  $N_{\text{cit}}/N_{\text{aut}}$  is dominated by the total number of citations to single-authored papers. It is smaller for women notably because:

- a) Women write relatively fewer single-authored papers than men.
- b) Women's single-authored papers are cited less than single-authored papers by men.

Citations to women's single-authored papers are less than those to men's papers because men cite women less. This has a large impact because there are so many more men than women.

# Interpretation?

We cannot “blind” author names on papers that were published and cited in the past.

Bibliometric analyses therefore can neither reveal biases against women nor prove that such biases do not exist. It documents what researchers do in their publications, but does not tell us why they do what they do.



# What can bibliography do for science?

Scientists presently use over-simplified measures: the h-index, total number of citations, the journal impact factor – and little else.

This:

- leads to perverse incentives
- makes those measures susceptible to gaming and eventually useless
- streamlines research practices/reduces diversity

# We know we shouldn't use those measures...

... but do it anyway. That's because scientists and HE admin likewise need ways to measure scientific impact quickly and easily.

Of course such measures can not replace a qualitative, in-depth assessment. But they are necessary and this need will not go away.

Even the computation of simple measures like the ones I discussed here is **prohibitively time-consuming** for individual scientists or university personnel.

# If you measure, measure right

SciMeter.org: A web-interface for research evaluation that is

- **Customizable and adaptive**  
Define your own measure for research impact
- **Transparent and reproducible**  
Both data and algorithms are documented
- **For Scientists, not against them**  
Designed to help scientists make fast assessments

That's it. Thanks for listening.