Technical University Of Denmark

Building a multi-year gendered dataset on attendance at physics scientific conferences

Supervisors:
Laura Alessandretti
Jonas Lybker Juul

Ellen Winther Mortensen, s184313

December 14th 2023

Contents

1 Introduction

People have been fighting for equality for many decades, but in recent years equality has become more important in the industry. Especially gender equality gets a lot of attention, and most organizations want to show that they are working towards a more equal organization. However, one thing is what organizations communicate about their equality programs, another is how equality actually is within the organization. This project aims to investigate gender equality at physics scientific conferences. American Physical Society(APS) is an organization that is working towards advancing knowledge of physics, and hosting several annual conferences. They have the programs 'Women in Physics'[1], 'Minorities in Physics'[2], and 'LGBT+ Physicists'[3], so equality is clearly on their minds. Therefore it's interesting to investigate equality for the American Physical Society. This project seeks to build a dataset of attendance at physics scientific conferences over many years, which later on hopefully can give insights into gender equality since the field of gender equality is filled with many opinions and a lot of misinformation. Gender equality will be investigated by studying the invited speakers at the March Meeting hosted by APS.

2 Scope

The scope of this project is to only investigate gender equality, even though there are many other aspects of equality. Furthermore, for simplicity, only two genders are assumed. This also means that we do not take into account how the attendees identify, but gender is estimated purely on their first name. This is heavily simplified but has been necessary as a starting point. The project is carried out with inspiration from a thesis[4] investigating gender equality within computer science. The main focus of the project is to build a dataset, that can later be used for a more thorough analysis of gender equality.

The investigation of gender equality of attendance at physics scientific conferences is limited to one conference; the APS March Meeting. Gender equality will be investigated by comparing invited speakers with the presenters. The invited speakers are researchers invited to present an abstract. Since the speakers need an invitation, this is a great measure of equality and opportunities across genders. The presenters are researchers presenting a scientific abstract. The presenters need to be an APS member and register for the meeting to present[5]. The presenters are therefore a sample of physics researchers. Although this might not be the case, it is assumed that the sample

and the population of physics researchers have the same distribution regarding gender. Therefore, the assumption is that gender equality would result in the same gender ratio for invited speakers as for presenters.

3 Methods

The project consists of web scraping, including data parsing, using an API to get more information about the researchers, and estimating the gender of the researchers. All these methods are described in this section, and the Python code for it is available on GitHub[6].

3.1 Webscraping

To investigate gender equality of attendance at physics scientific conferences, a dataset will be built. The dataset is built by web scraping information from the APS home page, where the Python library `Beautiful Soup` is used.

The starting point for the web scraping is an overview of the Meeting Archive on APS[7] which includes 1993-2023. Each year links to a list of meetings in that year. Each of the meetings in the list links to the page for that particular meeting. There are links for the March Meeting for 1994-2023 except for 1999. This leaves us with March Meetings for 29 years. The structure of the March Meeting page is not consistent over the years. Therefore, the first aim is to find a base page, that is the same for all years. This will make scraping information later on much easier. This was however not possible, because the homepage has changed during the last 30 years. Instead, we are working with two types of base pages. We call the first one 'Bulletin home', which shows all sections for the page of the meeting. The second type of base page is an overview of all sessions at the conference. Figure 1 shows the process of scraping the link for the base pages.

After getting the base page for all available years, the next step is to web scrape information about the sub-sessions. The method for this depends on whether the base page is the bulletin home or an overview of all sessions. Figure 2 shows the method of web scraping information about the sub-sessions.

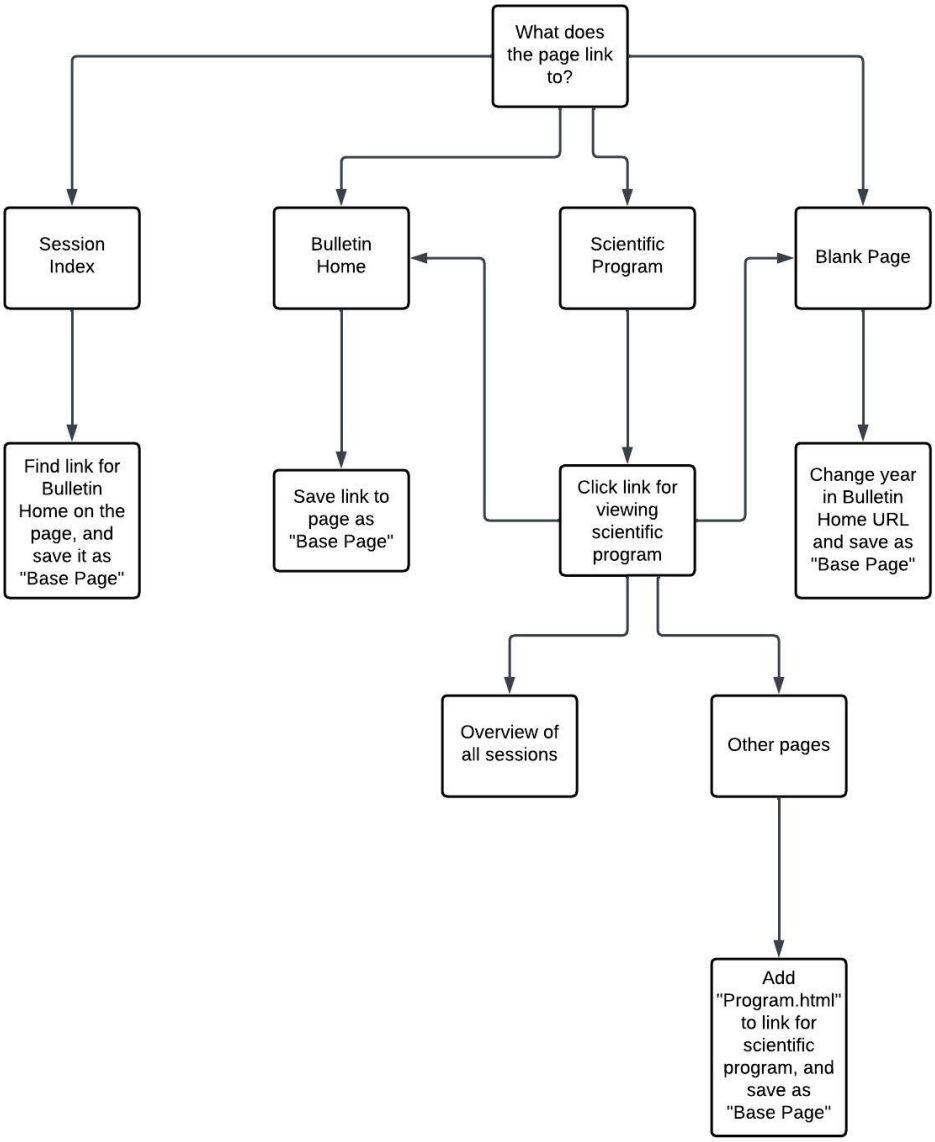


Figure 1: Method to web scrape Base Page. The link to the March meeting does either link to a page with Session Index, Bulletin Home, Scientific Program, or a blank page. The arrows show what to do depending on the case. Other pages is a page, that is none of the described but can look different depending on the year. Bulletin Home URL is the URL for Bulletin Home for 2023[8].

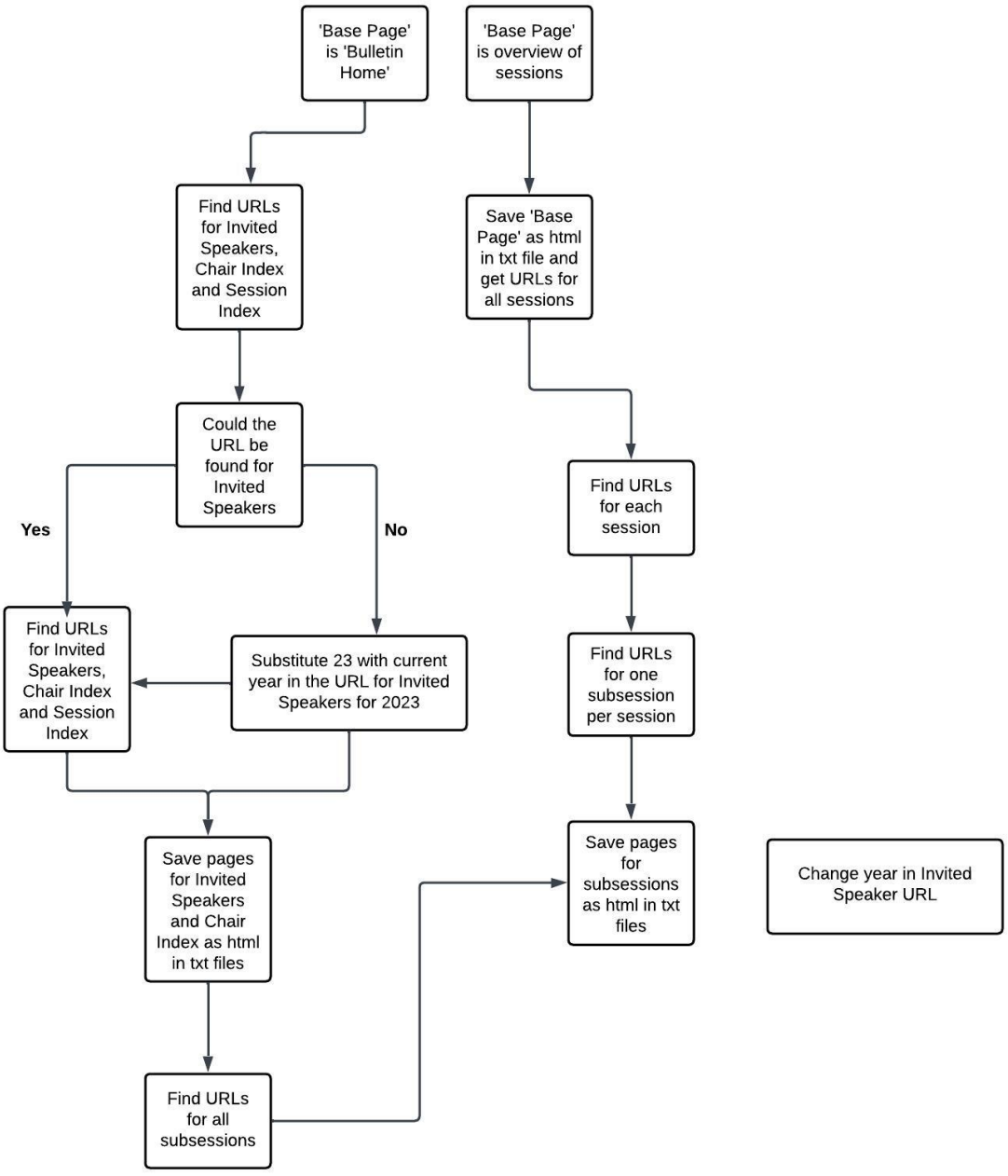


Figure 2: Method to web scrape information about sub-sessions. Only the last two digits of the year are used when substituting 23 with the current year. For the years, where the base page is showing 'Bulletin Home', three types of files are saved; Invited Speakers, Chair Index, and sub-sessions. For both Invited Speakers and Chair Index is saved one file per year. Pages for sub-sessions contain all the given information for a sub-session. One file per sub-session is saved. For the years where the base pages are an overview of all sessions, the overview is saved. Furthermore, is saved one file of the sub-sessions for each session.

The web scraping for base pages showing 'Bulletin Home' is pretty straightforward, since we can just follow the links on the page.
The page for Invited Speakers is saved, which is a list of the invited speakers, the institution of the speaker, the session title, and the abstract title. The page for Session Index is saved which contains the name of the persons chairing the session, the institution of the chairing person, and the title of the session they are chairing. Lastly, the Session Index is saved which contains information about the name of the session, the sponsoring unit, names of who is chairing the session, the abstract title, the authors of the presented abstract, and the institutions of the authors. Sometimes the presenter is stated as well. The sponsoring units state which topic the session is within.
For the base pages that show an overview of all sessions, it is a bit more complicated, because there is no menu as there is on the 'Bulletin Home'. This overview contains information about the sponsoring unit, the title of the session, presenters, and the name of the person who is chairing the session and is saved as a file. All sub-sessions of a session link to the same page, which shows an overview of the sub-sessions. This means that if there are five sub-sessions, then we will have five links to the same page. To reduce computational time, only one sub-session overview page per session is saved. This page has information about the title of the session, the title of the sub-session, the session type, the speaker/authors of the abstract, and the institution of the speaker/authors.

During web scraping, the pages with needed information are downloaded instead of processed immediately after accessing them. There are two important reasons for this, 1) saving time since it turned out to be more effective and most importantly 2) the web scraping was very early in the project, which meant that the knowledge of what information was important was minimal. When downloading files, it allowed to web scrape, and later choose what information to use, since the information was already in the downloaded pages.

In the process of web scraping it was checked for several years, if the downloaded pages seemed correct when compared to the website. However, not everything was checked, so there could be errors, especially since we are working with HTML, which can look very different from year to year. Furthermore, it was found that some links did not work, why not all sessions could be downloaded.
There were some problems with web scraping data for all of 1994 and some of 1995. Therefore, it was chosen to remove all data for 1994 and 1995.

After downloading all necessary pages, the first step of the web scraping is

done. The next step is to parse the data from HTML to a more structured data format.

3.1.1 Data Parsing

After the web scraping was finalized, the next part was to parse the HTML to a more structured data format. The aim was to create a dataframe with the invited speakers and presenters, which could later be enriched with gender and additional information about the speakers and presenters.

In the data parsing, all necessary information from the files is collected in several dataframes. For the names, some parsing problems were faced, primarily due to non-English letters and references in the names. Names with references mean that there might be some letters or numbers in the names, which is not part of the name. It is crucial to parse the real name since otherwise, it will likely be impossible to get more information about the researchers using the API. The parsing problems were solved to some extent, but there are still some problems with names not being correct.

The project aims to build a dataset with invited speakers and presenters. For 2018-2023 the presenters were stated, however, for the rest of the years, only the authors were stated. Therefore, all authors are saved instead of only the presenter. The authors can, just as the presenters could, give insight into the gender ratio in physics. However, some of the authors might not be in the field. It will later be checked if all presenters are in the field of physics, but as a starting point, all authors are saved.

The data parsing ends the web scraping, and enriching the data with information about the researchers can begin.

3.2 API

OpenAlex is used to gather more information about the speakers and authors. OpenAlex is a free API with information about authors[9]. For the rest of the report author will reference to authors of the presented abstracts as well as the invited speakers, unless otherwise stated.

The API returns a lot of information, some of which are author count, display name, fields, paper count, and number of published papers shown by year. The API often finds data for several names for one author, even though it sometimes seems unlikely that it is the same author. The number of names found for one author is what the author count is. The API works by only returning the first 25 results, this can however be changed [10]. It was chosen

to use all results up to 25 for an author, even though it might not be the same person. If we only took the first result for the author, we might lose a lot of information.
The display name is not necessarily the name searched for, but the one displayed. This name will be saved as the API name for each author.
The API often returns many fields, why physics is assigned as the author's field if that is one of the returned fields. This enables us to later only investigate authors within physics. The paper count is the number of published papers in total.

For every author, the year they published their first paper is derived. This is derived from the overview of the number of published papers shown by years. However, it turned out that this overview only showed the number of published papers from 2012 onwards.
The cumulative paper count is calculated based on the year of the conference and the overview of the number of published papers shown by year. The cumulative paper count is the number of published articles up until the year before the conference. Since the March Meeting is in the first quarter of the year, only paper published before the conference year is part of the cumulative paper count.
The seniority is calculated as well, which is calculated as year of conference − year of first published paper, which gives information about how many years an author has been in the field, at the time of attending the conference. Lastly, the productivity is calculated, which is calculated as $\frac{CumulativePaperCount}{Seniority}$, meaning that it is their productivity when attending the conference. The productivity is the number of published papers per year since they published their first paper. It should be noted, that since there are problems with showing the number of published papers for each year they have been in the field, the seniority and productivity will not be correct for authors publishing before 2012.

The API is used before estimating the gender because it ensures that all rows in the data are authors, and works as a way to clean the data. Furthermore, it allows us to get the first name, instead of having the first letter as the first name. However, this is not always the case, since the API sometimes returns one letter as the first name. Furthermore, the API sometimes estimates the name only based on the first letter of the name.
Each author has a name before the API; the original name, which is the one scraped, and a name after using the API, which is the display name from the API. Sometimes the name before and after the API is the same, and sometimes it differs. The author's first name is found by taking the longest first name of the original name and the display name from the API. In that

way, we use the full first name if it is available instead of using one letter as the first name. As a final cleaning step, authors with a first name including a dot, are removed. Authors with a dot in their first name are typically because their first name only consists of the first letter in their first name, due to the API not displaying the entire first name. The problem with this is that gender will most likely be estimated based on this, and it is likely the gender will be wrong. Therefore they are deleted instead.

3.3 Gender estimation

Now that the names have been cleaned to some extent, and we are only left with authors, the gender will be estimated. This will work as a last cleaning step since only first names also present in a gender dataset, will be saved.

The gender estimation is performed using "The World Gender Name Dictionary", WGND 2.0 from Harvard Dataverse[11]. The only information needed to estimate the gender is the first name and country code. Country code is needed since the gender of some names differs across countries. One of the attributes in gender data is `weight`, which is the probability of a name belonging to a given gender with the given country code. The country codes were not available through the API, so as a starting point, US was used as the country code. The gender was estimated by assigning the gender with the highest probability for that name. It was not possible to find all the authors with US as the country code. Therefore, the gender was assigned for the rest of the authors by assigning the gender with the highest average probability across all countries except the US. There were 25379 authors whose gender was not estimated either due to their first name not being in the gender data, or the gender being assigned to "?". Only authors with gender assigned to either male or female are used for the rest of the project.

4 Results

In this section will the results of the web scraping, API, and gender estimation be shown. Furthermore, will be shown plots generated to analyze gender equality.

4.1 Webscraping

The result of the web scraping is two dataframes. An overview of the dataframes is shown in this section.

One dataframe is with data regarding the chair index, which is shown in table 1. The second dataframe contains the invited speakers and the authors of the presented abstracts, which is shown in table 2.

Attribute	Type	Description
Year	Integer	Year of conference
Name	String	Name of person chairing a session
Institution	String	Institution of the chair person.

Table 1: Overview of attributes for Chair index

Attribute	Type	Description
Year	Integer	Year of conference
Name	String	Name of author
Institution	String	Institution of author
Title	String	Title of abstract
Division	String	Specialized field (sponsoring unit[12])
SessionType	String	Invited speaker or author of presented abstract
SessionIndex	Decimal	Index of session

Table 2: Overview of attributes for invited speakers and author of presented abstracts.

After the web scraping the data consists of 2414108 rows with researchers, with 227030 unique researchers. However, it should be noticed that some rows are not researchers, since there have been some issues in the parsing. Furthermore, some researchers might change their names, so there might be even more unique researchers in the data.

4.2 API

The result of using the API is one dataframe, which is the dataframe in table 2 with some added attributes. Table 3 shows an overview of the added attributes which consists of the authors found by the API.

After using the API and removing authors with a dot in their first name, we ended up with 1695315 authors, with 183272 unique authors. There might be even more unique authors if their names are not written the same way each time.

Attribute	Type	Description
FileID	Integer	FileID for API file
OriginalName	String	Name of the author from the web scraping
API_Name	String	Name of author returned by the API
MaxPaperCount	Integer	Total number of published papers
Field	String	Field of study
FirstPaperYear	String	Year of the first published paper
AuthorCount	Integer	Number of results from API when searching on the author
FirstName	String	First name of the author
Seniority	Integer	Number of years in the field at the time of the conference
CumulativePaperCount	Integer	Number of published papers at the time of the conference
Productivity	Decimal	Productivity at the time of the conference

Table 3: Overview of attributes for invited speakers and author of presented abstracts, after using the API.

4.3 Gender estimation

The result of the gender estimation is one dataframe, which is the dataframe from table3 with one added attribute. Table 4 shows an overview of the added attribute, which consists of the authors where gender was estimated.

Attribute	Type	Description
Gender	String	F for female and M for male

Table 4: Overview of attributes for invited speakers and author of presented abstracts, after estimating the gender.

After estimating the gender, we are left with 1467456 authors, with 157893 unique authors. Again, there might be even more unique authors. Lastly, authors with other fields than physics are removed. This leaves us with 1457833 authors, which 155730 of them are unique authors. This is the final dataset, which will be visualized to give an overview of the data and get some insights into gender equality.

4.4 Summary statistics and visualization

This section focuses on visualizing the data and its summary statistics regarding gender equality.

Figure 3 shows the number of results for each searched author from the API. As mentioned earlier, often the API finds several results for one author, which this plot shows. However, the plot shows that the mode is at 0, which is 1 result on an ordinary scale. The plots show that many authors have between 3 and 148 results on an ordinary scale, so clearly, a lot of authors have unrealistically many results.

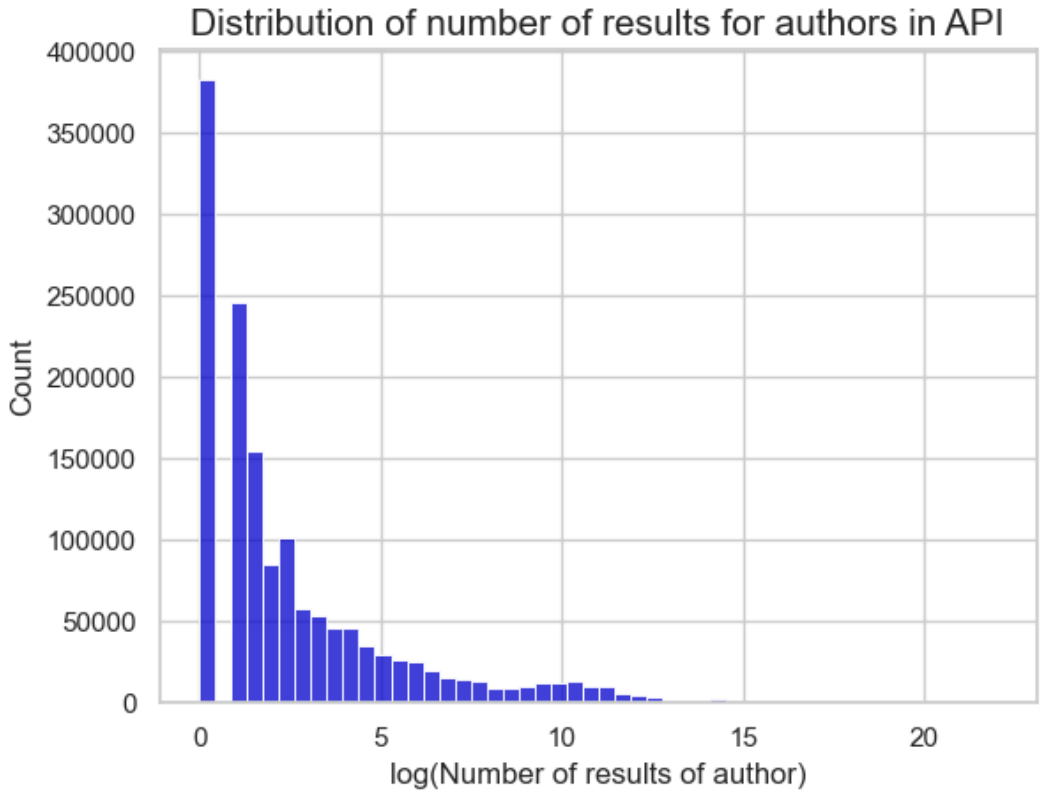


Figure 3: The figure shows the distribution of the number of results for authors returned by the API. The number of results is shown on log scale.

Figure 4 shows the number of invited speakers for each year. In general is observed an increasing trend until 2003, where it starts to decrease. The number of invited speakers increased again from 2013. However, especially 2002 which is very high and 2010 which is very low seem like outliers, why this should be investigated in the future. 1996 is also very low, and it could indicate that only very little data was accomplished to web scrape.

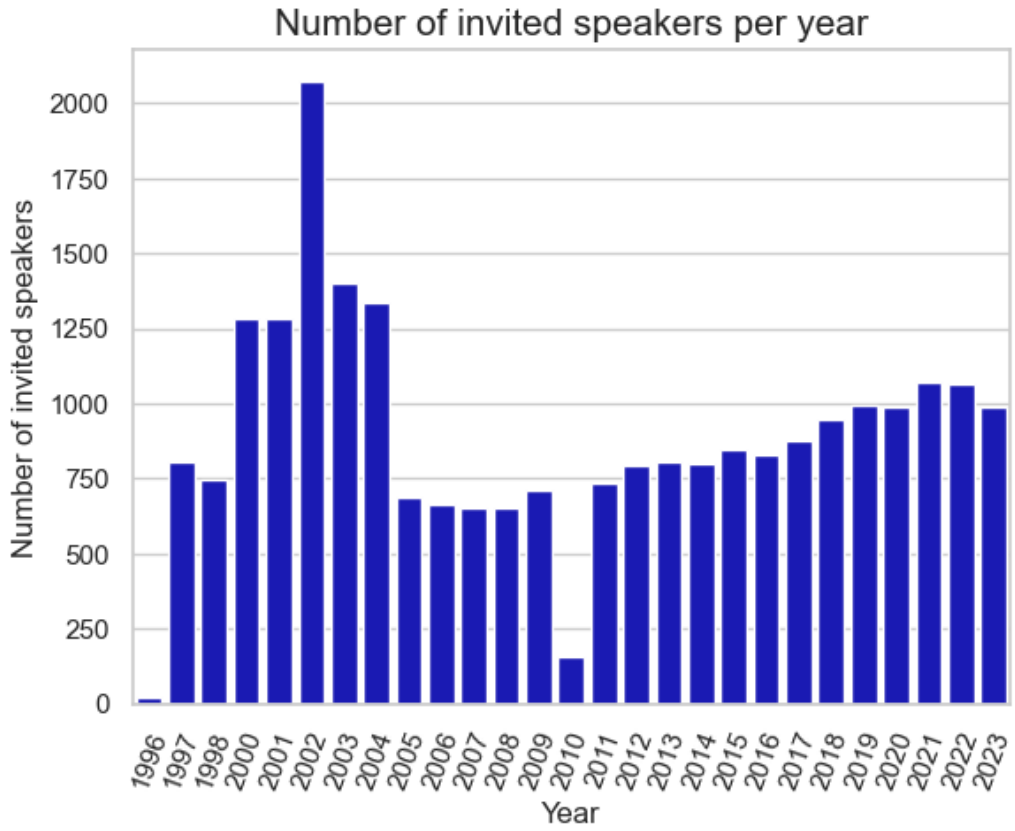


Figure 4: The figure shows the number of invited speakers per year.

Figure 5 shows the number of authors of presented abstracts for each year. In general is observed an increasing trend until 2004, where it starts to decrease, and from 2005 the trend is again increasing. Again is almost no data for 1996. For this plot, it is 2003 which is very high, and not 2002 as it was for the invited speakers. The number of authors for 2003 is more than 300000 higher than any other year, which indicates that something is off. Therefore this should be investigated further.

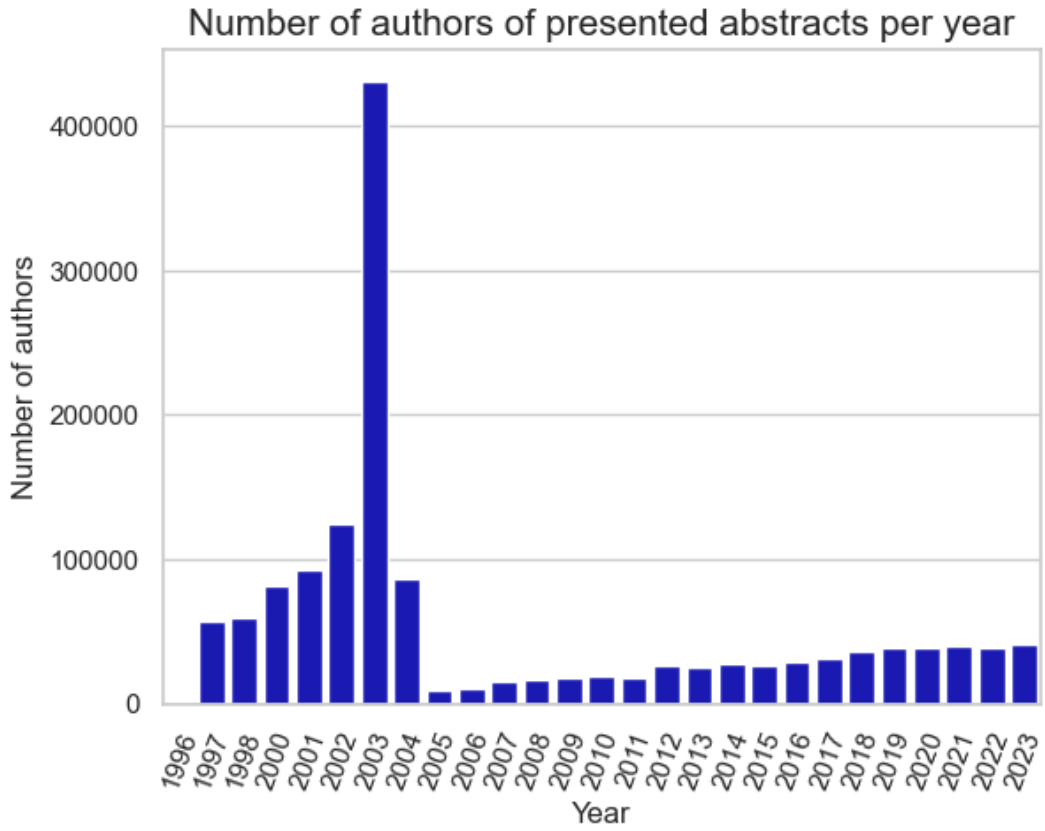


Figure 5: The figure shows the number of authors of presented abstracts for each year.

Figure 6 shows the gender ratio for authors of presented abstracts. The gender ratio is quite stable all years around 0.2. To have equality between genders, we would say that the gender ratio for invited speakers should be around 0.2 as well. Figure 7 shows the gender ratio for invited speakers, where the female ratio is increasing from 1996 to 2023. More than that, the female ratio is above 0.2 from 2016. This suggests that women do have the same opportunities as men, or even better opportunities.

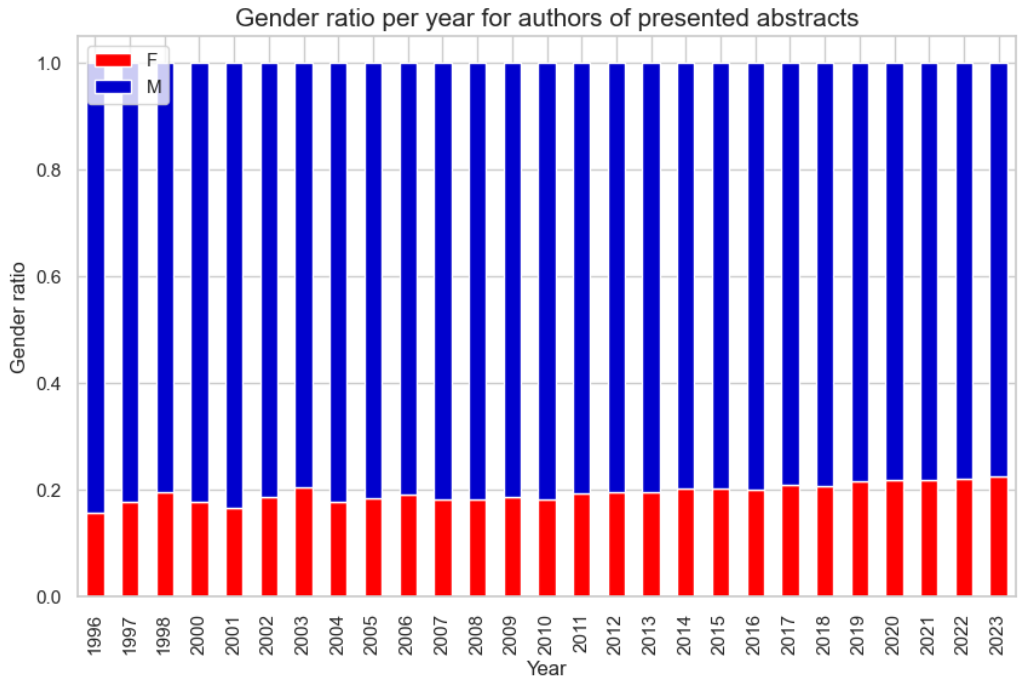


Figure 6: The figure shows the gender ratio for the authors of the presented abstracts for each year. The gender ratio is shown such that the sum of the ratio for male and female sums to 1 each year.

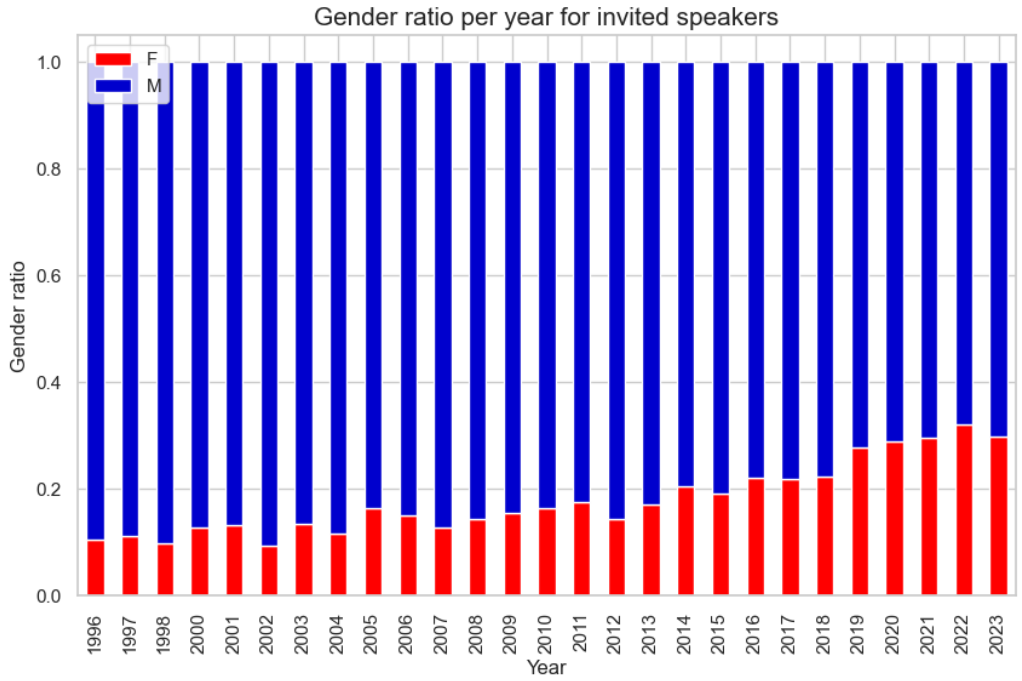


Figure 7: The figure shows the gender ratio for the invited speakers each year. The sum of the ratio for male and female sums to 1 each year.

Figure 8 shows the number of female and male invited speakers, and figure 9 shows the number of authors for presented abstracts with female and male indicated as well. Since we know there are some problems with the data, it is great to see that the distribution of the number of invited speakers for males and females is quite similar. The same can be said for the distribution of the number of authors of the presented abstracts. However, it should be noted that the number of male invited speakers was very high in 2002 and that the number of invited speakers for males and females was low in 2010. This corresponds fine to what we saw in figure 4. Figure 9 shows that the number of authors of the presented abstract is high both for males and females in 2003, but is especially high for males. This plot corresponds fine with what we saw in figure 5.

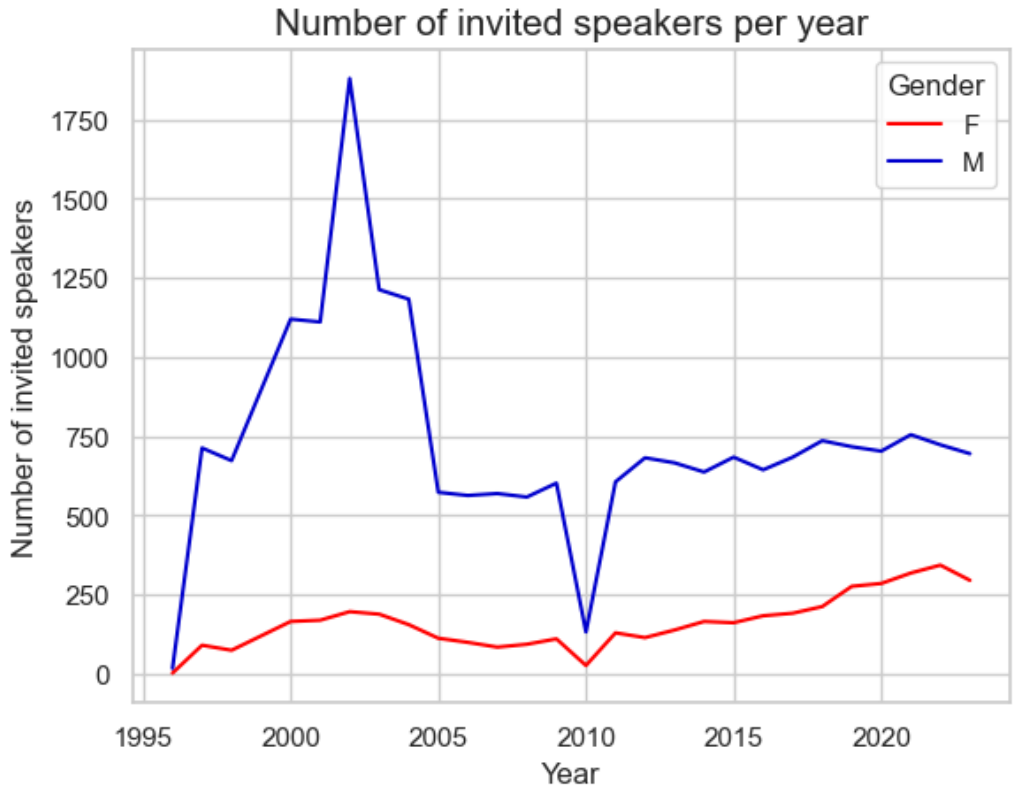


Figure 8: The figure shows the number of invited speakers per year. The blue graph shows the number of male invited speakers, and the red graph shows the number of female invited speakers.

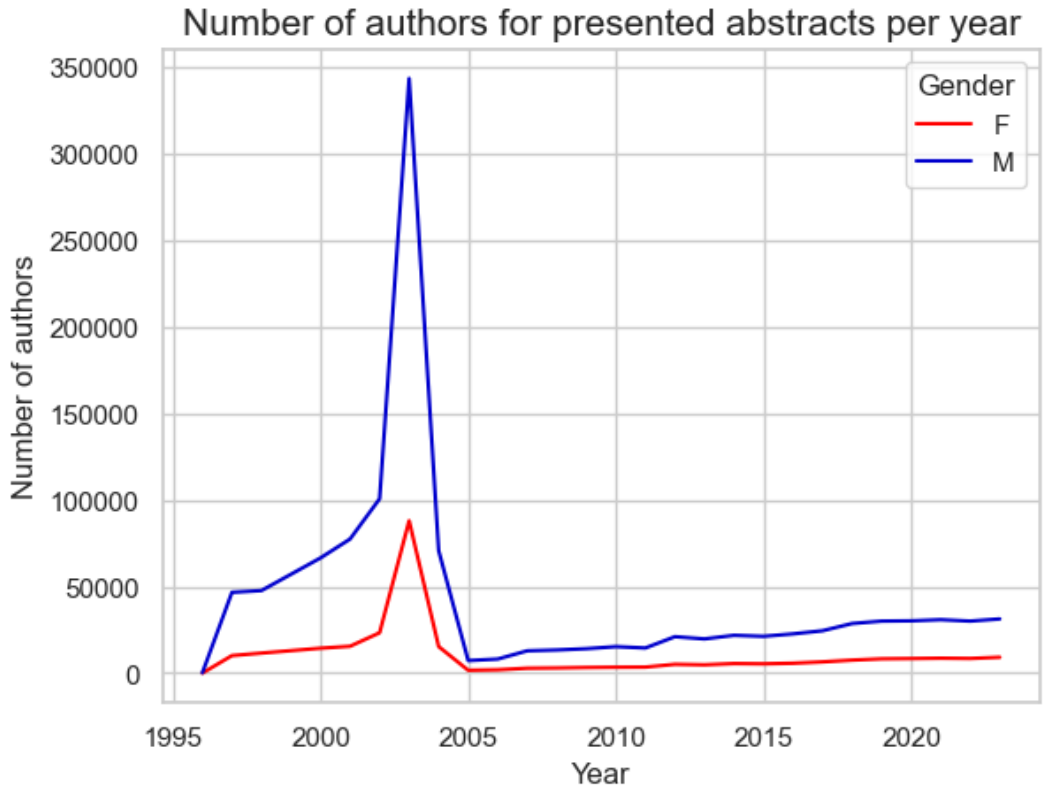


Figure 9: The figure shows the number of authors of the presented abstracts. The blue graph shows the number of male authors, and the red graph shows the number of female authors.

Figure 10 shows the ratio of females per year for invited speakers and authors of the presented abstracts. The two graphs intersect around 2015, and from 2016 the female ratio is larger for the invited speakers than for the authors of the presented abstracts, and it keeps increasing. It goes from 0.1 to 0.3 in 30 years. From 2016 until 2022 the ratio of female invited speakers increases faster than earlier.

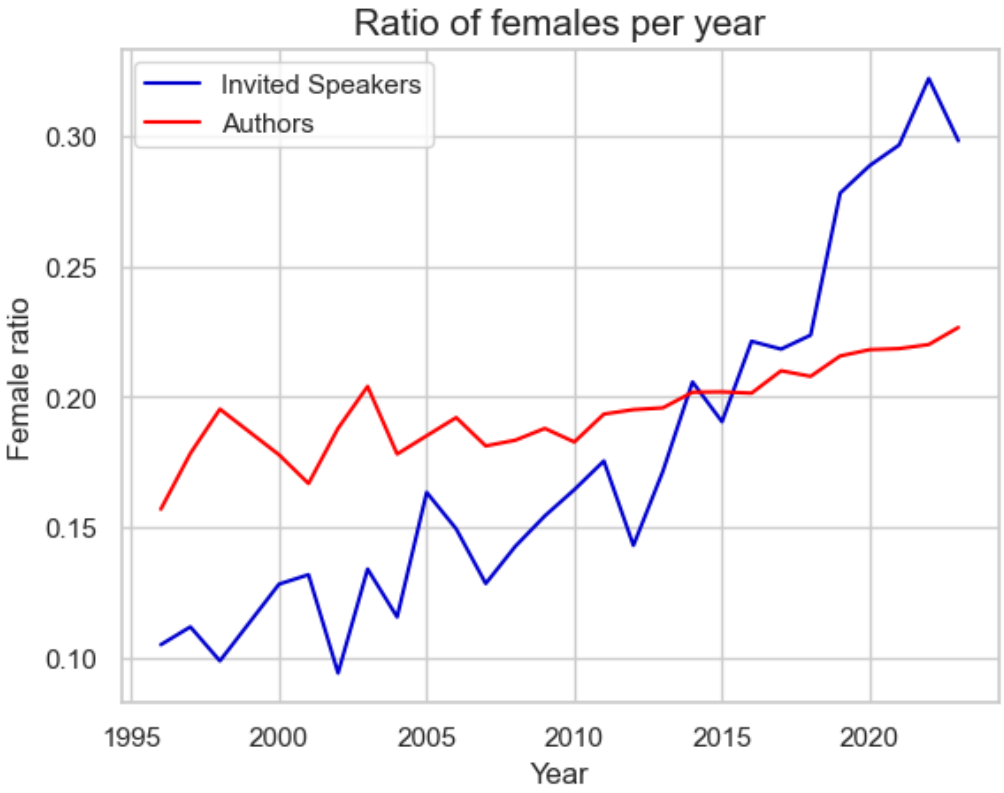


Figure 10: The figure shows the ratio of females for each year. The blue graph shows the female ratio of invited speakers, and the red graph shows the female ratio for authors.

Figure 11 shows boxplots of the number of published papers in total. The plot is only shown for authors with one as author count from the API since it would otherwise give unrealistically numbers. For example, did one author publish 65866 papers in total according to the API. The plot indicates that men publish more papers than women. This is for invited speakers as well as authors of presented abstracts. However, notice that the invited speakers tend to publish more than the authors of presented abstracts.

Figure 12 shows the distribution of seniority for men and women, for invited speakers and authors of the presented abstracts. It seems the male invited speakers and male authors have the same seniority. Female invited speakers have higher seniority than when being authors of presented abstracts. They also have lower seniority than the men when comparing authors of presented abstracts. However, when comparing invited speakers, females have higher seniority than males, at the time of the conference. This could indicate, that women have to fight harder or be more qualified to be invited. However, it should be kept in mind, that this is not the entire data, and that there were

some problems calculating seniority due to the API.

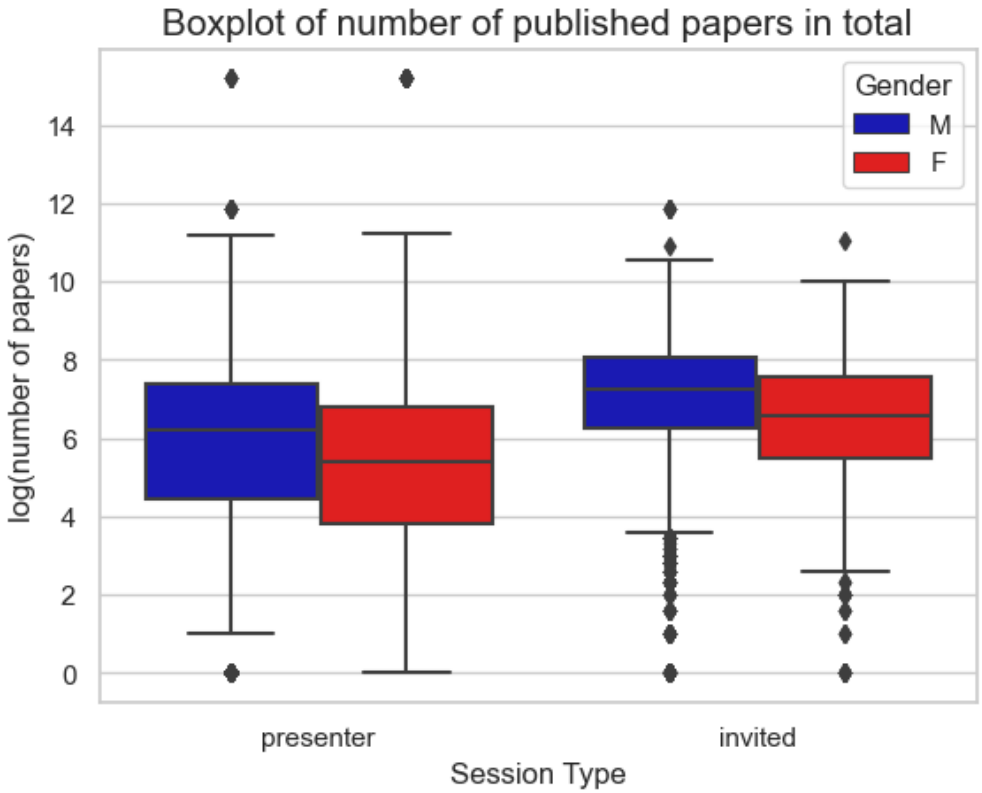


Figure 11: The figure shows a boxplot of the number of published papers in total. The blue boxplots show the distribution for the males and the red boxplots show the females. The boxplot for the presenter is the distribution of the authors of the presented papers, whereas invited is for the invited speakers. The y-axis shows log(number of published papers). Note that this is for the total number of published papers, and does not take the year of the conference into account.

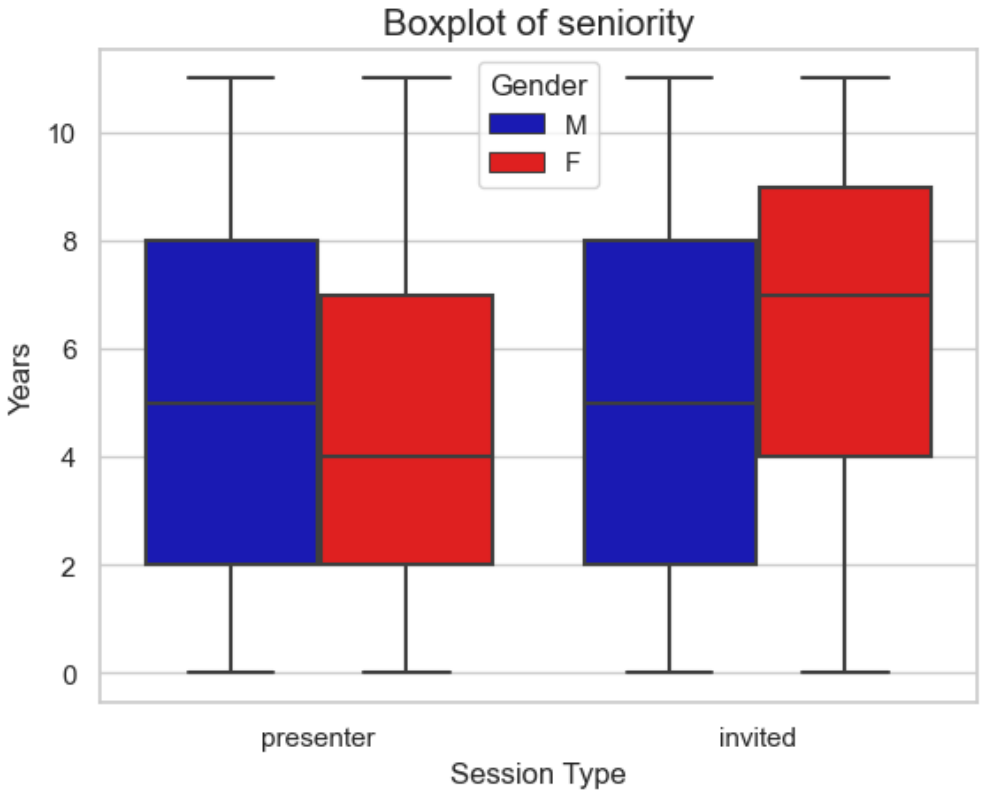


Figure 12: The figure shows boxplots of the seniority, so the number of years in research. The blue boxplots show the distribution for the males and the red boxplots show the females. The boxplot for the presenter is the distribution of the authors of the presented papers, whereas invited is for the invited speakers. The boxplot is shown only for authors with only one result from the API, and only for authors with positive seniority.

Figure 13 shows boxplots of the distribution of productivity. This suggests that women are less productive than men, and that invited speakers might be a little more productive. However there are a lot of outliers, so this should be investigated further. Just as with the seniority, it should be kept in mind that the boxplots do not show all the data and that it is only shown for positive seniority. One reason for this might be that women tend to spend longer time on maternity leave, however, this can vary a lot between nationalities.

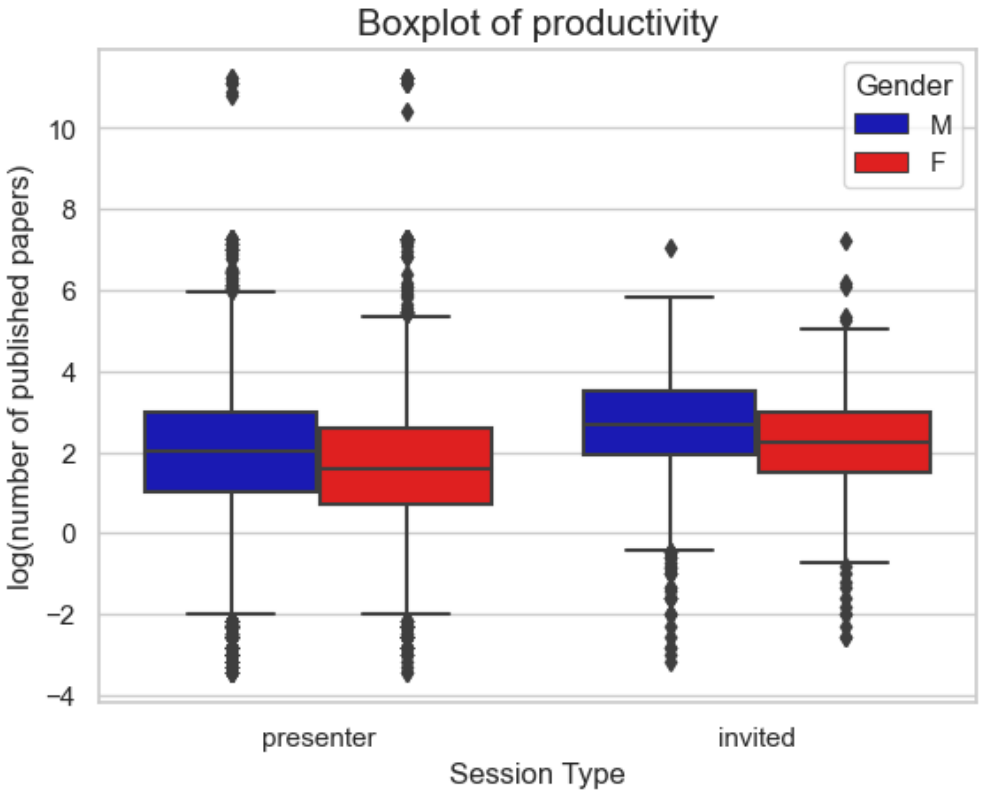
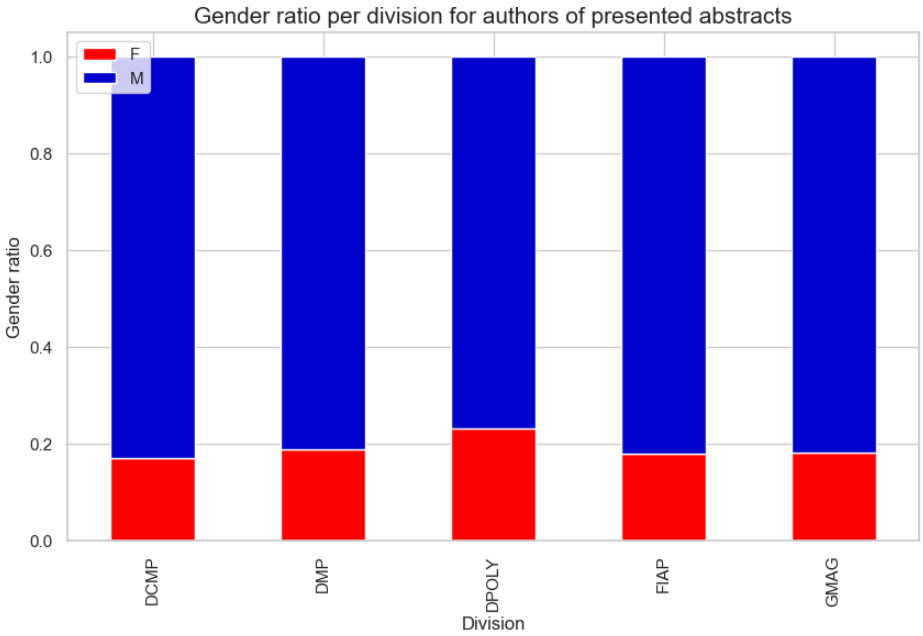
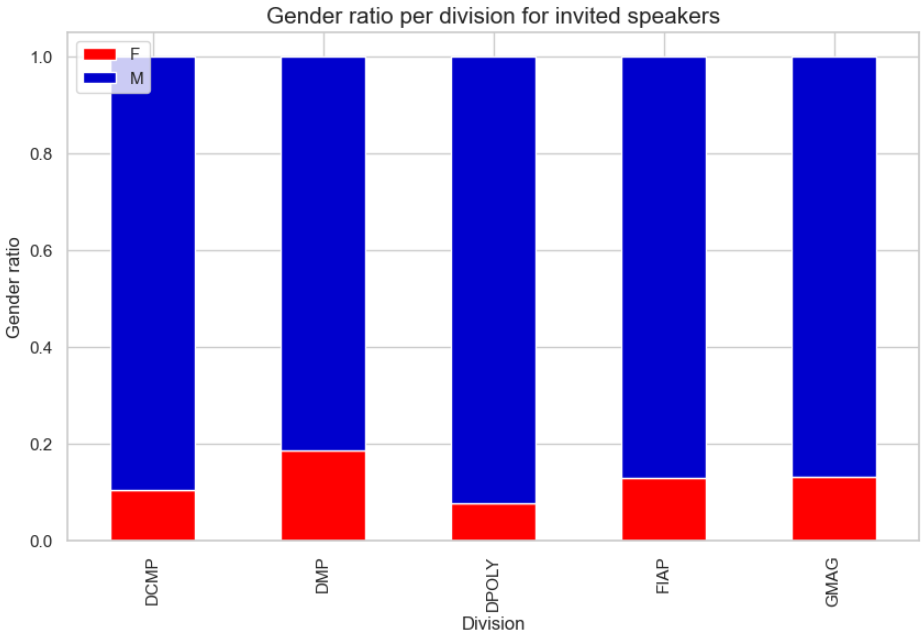


Figure 13: The figure shows a boxplot of the productivity, so the number of published papers per year they have been in research. The blue boxplots show the distribution for the males and the red boxplots show the females. The boxplot for the presenter is the distribution of the authors of the presented papers, whereas invited is for the invited speakers. The y-axis shows log(number of published papers). The boxplot is shown only for authors with only one result from the API, and only for authors with positive seniority.

Figure 14 shows the gender ratio for authors of presented abstracts and the invited speakers, for five divisions within APS. The five divisions are assumed to be the five largest divisions since these are found most times as the sponsoring unit. The plots show that the female ratio is around 0.2 for all five divisions for the authors of the presented abstract, whereas it is lower for all five divisions for invited speakers. This suggests that gender equality is not reached, which is the opposite of what figure 6 and figure 7 suggested. This indicates that gender equality differs across divisions and that other divisions must have a higher ratio of female invited speakers.



(a) Authors of presented abstracts



(b) Invited speakers

Figure 14: The plots show the gender ratio within the five largest divisions for (a) the authors of the presented abstracts and (b) the invited speakers. The divisions shown are DCMP: Condensed Matter Physics, DMP: Materials Physics, DPOLY: Polymer Physics, FIAP: Forum on Industrial & Applied Physics, GMAG: Topical Group on Magnetism and its Applications.

5 Discussion and conclusion

A dataset of 1457833 researchers enriched with publication data and gender, has been built. The data was cleaned to some extent, but the data could need more cleaning. By not cleaning the data properly, some names are deleted, and more importantly, some first names are estimated by only having the first letter of the first name, which makes it very likely to get misleading information.

Some issues were discovered about the API, such as estimating different authors as being the same. Furthermore, data about when papers are published, are only shown from 2012, making it impossible to calculate seniority and productivity correctly. Therefore, as part of the future work, another API should be used. If this API can also return the country code of the author, the gender can be estimated more precisely than now.

Despite some issues with the API and lack of cleaning, some clear trends are observed in the data. The analysis does however suggest opposite conclusions. It seems that women in general have the same opportunities as men regarding being chosen as invited speakers for the March meeting. More than that, it seems they have better opportunities than men. However, it seems that women need more seniority to be chosen as invited speakers than men. This suggests that equality might not be as great as some plots indicate and that women need to be more qualified to be chosen. This does however make more sense when you take into account that women seem to be less productive than men. That would mean that women in general need more years to reach the same number of published papers. Future work should include investigating seniority and productivity regarding the importance of these factors when inviting speakers. The female ratio within the five largest divisions showed that 0.2 seems to be a good estimate of the female ratio within physics since this ratio was observed for all five divisions for authors of the presented abstracts. It also showed that the female ratio was lower for invited speakers than for authors of the presented abstracts, meaning these divisions did not have as much gender equality as shown across all divisions. This suggests, that gender equality differs across divisions, why future work should also focus on establishing differences in gender equality across different fields within physics.

The analysis also showed outliers for the years 1996, 2002, 2003, and 2010, why these years should be further investigated in future work.
Information about chairing persons was web scraped. The data was however not analyzed, why future work could focus on this as well.

A large dataset has been built with attendees at physics scientific conferences. Female does in general have good opportunities, but this does not apply within all fields of physics. The influence of other factors needs to be investigated further before concluding anything, but interesting results were shown.

6 Literature

References

[1] *"Women in Physics"*. URL: https://www.aps.org/programs/women/index.cfm.

[2] *"Minorities in Physics"*. URL: https://www.aps.org/programs/minorities/index.cfm.

[3] *"LGBT+ Physicists"*. URL: https://www.aps.org/programs/lgbt/index.cfm.

[4] Marie Sjørslev Pedersen. "Evolution of Gender Diversity in Academia". Technical University of Denmark, 2023.

[5] *"Presenter Instructions Logistics"*. URL: https://march.aps.org/attendees-presenters/presenter-instructions-and-logistics.

[6] *"GenderGap-in-Physic"*. https://github.com/EllenWM/GenderGap-in-Physic.git.

[7] *"Meeting Archive"*. URL: https://www.aps.org/meetings/baps/index.cfm.

[8] *"Bulletin of the American Physical Society"*. URL: https://meetings.aps.org/Meeting/MAR23/Content/4348.

[9] *"A free index for the world's research"*. URL: https://openalex.org.

[10] *"Get lists of authors"*. URL: https://docs.openalex.org/api-entities/authors/get-lists-of-authors.

[11] *"World Gender Name Dictionary(WGND 2.0)"*. URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MSEGSJ.

[12] *"APS Units"*. URL: https://www.aps.org/membership/units/index.cfm.