M.Sc. Thesis
Master of Science in Engineering

DTU Compute
Department of Applied Mathematics and Computer Science

Gender and Language Differences in the Growth of Wikipedia

Investigating networks of scientists across language and time

Eline Brunke Aarhus
Karen Visby Østergaard

Kongens Lyngby
December 31, 2024

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Matematiktorvet
Building 303B
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

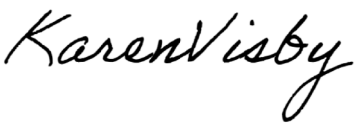
Abstract

Wikipedia is among the most visited websites worldwide, and being the most used online encyclopedia, the users expect neutral, correct and unbiased content. However, several studies have proved a gender bias in Wikipedia's linking structure, notability demands as well as article content, and found that Wikipedia, in many ways, has a bias against women. This thesis aims to investigate this bias by creating a network of Wikipedia biographies of scientists across multiple languages as well as investigating how these networks grow and if the patterns differ across gender and language. Analysing the network, looking at structure, gender distribution and growth, it is found that not only are the female scientists under-represented in the network, they are also less connected than the male scientists and they generally become part of the graph much later than the men. Community detection and text analysis showed that the scientists in our network are more likely to link to other scientists within the same fields as themselves. Statistical analysis also showed that new out-edges tend to be added to nodes that are located close to other nodes that recently gained out-edge, providing knowledge about how Wikipedia grows. Comparing the English network to the corresponding German, French and Spanish networks, many of the same tendencies were evident. These networks also have fewer and less connected women, but are also less densely connected than the English network. Despite this difference, the community structure across languages revealed that the networks also have many of the same features. And while the individual nodes do not grow the same across languages, the German, French and Spanish networks exhibit many of the same growth patterns as the English network. These findings combined suggest how Wikipedia grows as well as highlight the structure and bias within Wikipedia.

Preface

This master's thesis was prepared at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfilment of the requirements for acquiring a Master of Science and Engineering degree in Business Analytics.

Eline Brunke Aarhus
December 31, 2024



Karen Visby Østergaard
December 31, 2024

Acknowledgements

We would like to thank Sune Lehmann Jørgensen and Jonas Lybker Juul for all their help and guidance during the last 5 months. Thank you for sharing your great knowledge on networks, gender bias and the chaotic process of writing a master thesis - it has been a real pleasure to work with you both.

Contents

CHAPTER 1

Introduction

Since the founding of Wikipedia in 2001, the editor-driven online encyclopedia has grown to be one of the most visited websites worldwide [1] and a large source of information for users across languages and generations. Editors are able to create and add hyperlinks to and from articles, expanding a great network of articles and biographies.
With this extensive outreach, Wikipedia also has a large impact on the world and how it is understood. However, Wikipedia has been criticised for having a bias towards women, both in regards to the number of female editors[2], number of articles about women [3], and the content in the articles regarding women [4]. This bias and the behaviour of the growth and hyperlink structure are the main focus points of this thesis.

1.1 Background

1.1.1 Gender

Throughout this thesis, we will use *gender* as a definition when categorising data. Gender can be defined as socially constructed characteristics that mostly centre around binary definitions for men and women. Unlike biological sex, gender has changed over time and can also differ between societies. It builds on the concept of sex and the physiological differences between men and women but adds norms, behaviours, and roles associated with being a woman and a man. We will keep our focus on gender and not gender identity, which is related to gender but is more how a person feels and experiences gender on an individual level [5].

While the majority of people can be categorised by the two traditional gender definitions, non-binary people cannot be classified into these two categories. A study from 2018 approximated using the Wikidata Human Gender Indicator that around 0.0001 % of all Wikipedia biographies were about non-binary persons [6]. This is a very small demographic to consider and while the bias towards non-binary persons in Wikipedia is important, this thesis defines gender as binary being either male or female.

1.1.2 Gender bias in science

Our focus when scraping data will be Wikipedia biographies about scientists. In the science, technology, engineering and mathematics (STEM) fields, women have historically experienced less representation and been discriminated against in the educational system, the workforce and society at large. Gendered bias is a form of prejudice where men and women can be associated with different stereotypes or judged unfairly based on their gender. These biases can be explicit, meaning that a person is set in a belief and aware of the fact that they have a preference for one group over another. An example of this could be a person stating openly that they would rather work with a man than a woman because *a woman will not take their career as seriously as a man would*. Implicit bias is different, as this bias is not openly but rather unconsciously expressed. An example of this could be the automatic assumption that men are associated with both STEM and leadership and not assuming the same for women. Over time, there has been a reduction in the explicit bias towards women in STEM, the same cannot be said for implicit bias. This means that even though girls and women over time have taken up more space in

STEM, they still very much might be subjected to unfair prejudice [7].
While our focus in this thesis is on the bias and discrimination against women, other marginalised groups also experience discrimination, both in general and in the fields of STEM. In the U.S., the majority of people of colour working in STEM reported that they had experienced racial discrimination, and some expressed that the colour of their skin made an impact on how others view their work ethic and skills [8].

1.1.3 Historical bias

When scraping scientist biographies, we know that biographies have been written about scientists, from ancient to recent times. Historically, women have been under-represented in STEM fields, and while the number of female STEM workers is increasing [9], women are still not as common in STEM fields as men. For instance, the proportion of women in STEM jobs in the U.S. has increased from 8% in 1970 to 27% in 2019. In the same period the proportion of women in the American workforce went from 38% to 48% [10]. These numbers are, however, limited to the U.S. and STEM fields. Looking at another indicator, namely the proportion of female researchers worldwide, this number has increased slightly from 30.9% in 2011 to 31.7 % in 2021 worldwide. These numbers are only estimates and should be considered with some caution [11]. The numbers do, despite the uncertainty, show that the development of female researchers worldwide seems to have stagnated in the last 10 years. Though 31.7 % can seem far from equal, the problem might not be the proportion of female researchers worldwide, as the worldwide percentage of women in the workforce were estimated to be 40.1 % [12].

1.2 Scope

The scope of this master thesis is to investigate the growth of Wikipedia, unveil the differences between male and female scientists and look into whether Wikipedia has a bias towards female scientists.

To achieve this, we are going to work with the following research questions:

- How can we scrape revision data from Wikipedia articles about scientists and classify their gender?
- How can we construct networks across time and language versions?
- Are there differences in network structure across language and gender?
- Which growth patterns can be detected in the network across language and gender?

1.3 Outline of the thesis

This thesis consists of 6 other chapters, excluding this introduction chapter. Chapter 2 contains a literature review, going through existing knowledge on the subject. We will first cover Wikipedia and its hyperlink structure, then go through existing studies on gender bias in Wikipedia and what has been done to solve these problems. Chapter 3 covers how we scraped and modelled our data and detected the gender and birth year of the scientists from their biographies. Chapter 4 contains the methodology and tools as well as the results of the analysis of the English network. Chapter 5 covers the results of the analysis across the four selected languages. Finally, in Chapters 6 and 7, the findings of Chapters 3, 4 and 5 are presented and discussed, as well as future work and a conclusion of our thesis.

CHAPTER 2

Literature review

In this literature review, we aim to give a short introduction to Wikipedia, the existing knowledge on gender bias in Wikipedia, and what has been done to fight this bias.

2.1 Gender biases in Wikipedia

2.1.1 Wikipedia

Wikipedia is a free online user written encyclopedia founded in January 2001. The content is created by volunteer users all over the world and is administered by the non-profit organization Wikimedia Foundation [13]. Wikipedia's content is solely based on the work of volunteer users, so-called *editors*, and by January 2023 the English version of Wikipedia had more than 44 million registered editors [14]. In August 2024, Wikipedia was the seventh most visited website in the world, determined by the company SimilarWeb, and the most visited website in the category of dictionaries and encyclopedias [1]. Wikipedia thereby has a non debatable influence on the world and how their many readers view the world.

In this thesis, we aim to analyse Wikipedia data across the different language versions, and it is therefore relevant to find out which versions are the largest in regards to articles. If discounting the Cebuano version, the second largest Wikipedia with 6,116,880 articles but only 147 active editors [15], the English, German, French, Swedish, Dutch, Russian and Spanish Wikipedia are the largest measured by total number of articles ordered by size [16]. This does not directly correspond to what languages are spoken the most worldwide; here, the top 5 are English, Chinese (Mandarin), Hindi, Spanish and French [17].

Publishing an encyclopedia with such a wide range and readers all over the world comes with a responsibility to reflect the world as unbiased and true as possible. Being non-profit and based on volunteer work is not without consequences and several studies conclude that Wikipedia has a gender bias towards women[18]. This is seen in several different aspects. This literature review will cover some of the aspects in which gender bias is present in the following sections.

2.1.2 Number of biographies and deletion

As of August 2024, 19.91% of all biographies in the English Wikipedia were about women [3]. The reason for this much smaller amount could be that there historically are fewer famous/notable women and that Wikipedia not only has articles about famous people today but also historically important persons. However, studies have found that women who do have a Wikipedia biography are significantly more notable than men, measured by the number of Wikipedia language editions the given person appears in and the Google search volume of the person. In other words, the study concludes that it takes more for a woman to get a Wikipedia article than it does for a man [18].

A study from 2019 has investigated the number of biographies about American sociologists across gender and race and found that notability (measured by academic rank, length of career, H-index and

departmental reputation) only explains half of the likelihood of a female sociologist having a Wikipedia biography [19]. Another study shows that despite meeting Wikipedia's own criteria articles about women are more likely to be flagged as non-notable than articles about men that meet the same criteria [20].

Similar to the discrepancy between notability and the number of biographies, studies have also found a lack of coherence between notability and deletion of biographies. One of these studies, *"Too Soon" to count? How gender and race cloud notability considerations on Wikipedia* by Lemieux, M. E. et. al. found that while the deletion of white male academics' pages could be predicted by their online presence, this was not the case for female academics. The article, in particular, focuses on the label *Too soon*, which is a label editors can give a biography if they think the subject is not notable enough or the article lacks reliable sources [21]. The article finds that women's pages are more likely to get this *too soon*-label and that it is more often caused by the women's career stage rather than her achievements and online presence. The study concludes that female achievements count less than male achievements in order to have a Wikipedia article and avoid being nominated for deletion [22].

2.1.3 Hyperlink structure

Wikipedia is structured so that it is possible to link articles to other relevant articles using hyperlinks. This structure is very useful for understanding how Wikipedia articles relate to each other and how Wikipedia has developed over time. By representing Wikipedia as a social network with biographies as nodes and hyperlinks as edges, network theory has been used to detect gender biases in Wikipedia before. Using this approach, it has been found that female biographies are more likely to link to female biographies, and male biographies are more likely to link to male biographies. However, female biographies are more likely to link to male biographies than vice versa [4]. This is seconded by another study that concludes that "There are structural differences in terms of meta-data and hyperlinks, which have consequences for information-seeking activities" [18]. A third study used the term PageRank to measure node centrality based on network connectivity and found that male biographies were disproportionately more central than female biographies. This study created 5 different unbiased graphs with characteristics of the original network, so-called null-models, and compared these to the observed model. None of the null models showed any bias in link proportions but for the observed model female biographies tend to link to other female biographies. In other words men in the observed graph are significant less likely to link to women than in the 5 different null model, which indicated that men in the Wikipedia network are disproportionately more central than women [23].

Since Wikipedia articles aren't "locked" by the time they are created, it is likely that a page has had changes committed at different times from when the article was created up until now. These different versions of an article over time are also known as *revisions*, and all of them are stored and can be viewed in the article. It is possible for links to other biographies to be added in one revision of an article and removed in a later version. An example of this is the article for ancient mathematician Yusuf al-Khuri, which, on the 30th of January 2018 at 15:37, had a box added to the article that contained links to many other pre-17th century mathematicians in the medieval Islamic world. However, 6 minutes later, another revision of the article was made that this time, hid the box and, with that, removed the links. These rapid revisions can be explained by the fact that it is other users who are governing which editions are "true" facts about the topic of the biography. In theory, users are allowed to add anything they want to a Wikipedia article, but in reality, this is not the case, as other users are able to challenge revisions and re-edit links and text from the article that they deem irrelevant or not backed up.[24]

2.1.4 Differences between language versions

As of September 2024, Wikipedia had 332 active language versions according to WikiMedia [25]. Having this many different versions, the gender bias present in each version is possibly different. The previously mentioned study *Women through the glass ceiling: gender asymmetries in Wikipedia* found that the German version of Wikipedia has the lowest fraction of female biographies (13.2 %), while the Korean version has the largest (22.5 %). The German version is, however, remarkably larger than the Korean version (102,233 vs. 15,921 biographies) [18].

In a study by Young-Ho Eom et al. the above-mentioned term PageRank probability, a measure of how much other articles link to a given article, is used to quantify the differences in gender bias across different language versions. This study found that looking at the top 100 historical figures of each language according to PageRank, all language versions had a strong male bias. The authors do, however, acknowledge that this bias is not only a question about bias in Wikipedia but also a result of a history where women have had a smaller chance of becoming a historical person than men. Despite an overall biased tendency, the study was able to detect differences across language versions. Looking at PageRank it is found that a larger proportion of the top 100 historical figures in the Thai, Hindi, Swedish and Hebrew are women than the average across all languages. Finding the top 100 articles using 2D PageRank, a PageRank measure that besides measuring the number of ingoing edges also takes the number of outgoing edges into account, it was found that the English version of Wikipedia has a larger proportion of female figures than the average whereas the number for the German Wikipedia was less than average [26].

2.1.5 Editors bias

Being a volunteer editor-based website, the content of Wikipedia is a product of the many editors who daily contribute to writing, editing and fact-checking the articles. The demography of Wikipedia editors does, however, not reflect the world they are writing about. Since not all editors state their gender, it is difficult to find the exact gender distributions of the Wikipedia editors, but a 2011 study showed that 91% of all Wikipedia editors across all languages were men [2]. In 2015 the Wikimedia Foundation founder Jimmy Wales admitted that the foundation did not reach their goal of 25% female editors by that year[27] and by 2018, the gender distribution did not seem to have shifted as a user survey found that 90% of editors at that time identified as male, 8.8% as female, and 1% as other [28].

Not only are there significantly fewer female editors, but the women who write and edit articles are doing it significantly less than the male editors. The previously mentioned 2011 study by Wikimedia concluded that female editors were more likely to edit 1-50 articles during their lifetime than male editors, where male editors were more likely to make 10,000+ edits during their lifetime than female editors [2]. This conclusion is backed by another study that shows that even though women accounted for 16.1% of the new editors in 2009, these specific new editors only did 9.0% of all edits made by the group of new editors.
The same study found that the gender gap for editors does not seem to shrink over time, and for editors with science as their area of interest, the percentage of female editors is significantly lower than for other areas. In fact, 5.2 % of science editors were women compared to 7.6 % for all types of editors (2008 numbers). The study concludes that there is a gender gap between the areas male and female editors focus on, where female editors tend to focus more on People and Arts, male editors focus more on Geography and Science [29]. This is also supported by an article from The Guardian (2014) that, as an example, found that the Wikipedia list of female porn stars had been edited more than 3,000 times and was very well structured compared to the Wikipedia list of female writers that was very poorly structured [30].

This combination of a gender gap between editors and within the content types could potentially result in an under-representation of the subjects female readers are interested in. Furthermore, the lack

of female editors in the science subject could cause a lack of representation of female scientists. This lack of representation is what we will also investigate further in this thesis.

2.1.6 Difference in language used to describe men and women

Looking at the words used to describe male vs female articles Wagner et al. found that looking at 6 different language versions of Wikipedia, all 6 language versions suffered from a language bias towards women. They divided the words of an article into 4 categories: Gender, Relationship, Family and Other. **Gender** counts words like 'man', 'woman', 'mrs', 'lady' and so on, **Relationship** counts 'married', 'divorced' and similar relationship-related words, **Family** is the number of words like 'kids', 'mother' and so on. Finally **Other** is the count of all other words. The study found that the female articles had 23%-32% words from the three categories (variation between the 6 tested languages), whereas male articles had 0%-4% in comparison. The most biased in this context appeared to be Russian with 32% and the least (but still remarkably biased) German with 23,1% words from the three categories in the female articles [4]. These results emphasize that Wikipedia is not only biased in terms of the number and connectivity of female articles, but the language used to describe women is also remarkably different from the language used to describe men.

2.1.7 Initiatives to fight gender bias

There is a broad understanding across literature and studies that concludes that Wikipedia indeed is biased against women in several different aspects. The problem have not only been identified, several initiatives to fight this imbalance has been tried. As previously mentioned, the organization behind Wikipedia, Wikimedia, had a goal of increasing the number of female biographies to 25% by 2015 [27], which was still not reached by 2018 [28], but since then, it has not been possible to find any defined goals from the foundation itself.

To increase the number of female representation in Wikipedia several so-called Edit-a-thons with gender representation as their main focus have been hosted. Edit-a-thons are gatherings, online or in person, where the hosts teach the participants how to edit and create Wikipedia articles, and the participants create and edit articles within a certain topic, in this case, notable women of specific fields[31]. The group *500 women Scientists* has hosted several Edit-a-thons in order to increase the representation of female scientists in Wikipedia [32]. As of 3rd of November 2024, the group's Wikipedia Campaign, consisting of 890 editors, has created 355 articles and added 62 links between articles. [33]. Another group of volunteer editors called *Women in Red* are writing articles about notable women who are missing a Wikipedia biography. The project started in 2021 and is named after the red links that occur in Wikipedia when an article has not yet been created. This project is, however, met with resistance from the remaining Wikipedia editors, who often flag the articles as non-notable or nominate them for deletion [20].

As previously mentioned it is difficult to find up-to-date numbers on the gender balance of Wikipedia editors so whether all of these initiatives has helped significantly is difficult to say. Despite thorough research on the subject the available literature does not seem to have been able to detect a positive development in the Wikipedia Gender Bias.

2.2 Using network theory to detect gender biases in Wikipedia

The aim of this study is to investigate if the gender biases seen in the mentioned aspects also can be seen when analysing a network of scientists over time. We have found that female editors are less likely to edit articles about science, and our hypothesis is that this, as well as the other previously mentioned aspects, will affect how female scientists link to each other. We also aim to investigate if it is possible

to see a difference in how these networks grew in the first years of Wikipedia compared to today to see if the focus on gender biases has led to a change.

CHAPTER 3
Data and Preparation

This chapter will uncover our data preparation. It will include how we scraped Wikipedia data for four different languages in the time period 2001-2024, classified the gender and estimated the birth year of the biographies. We will furthermore explain how we structured the revisions and the state of the final data going into the analysis. Lastly, this chapter includes results on counts of gender-related words and distributions of birth years.[1]

3.1 Data Scraping

3.1.1 Scope

For this project, we aim to create a network of Wikipedia biographies that link to each other according to the hyperlink structure in Wikipedia articles. Since we are researching the gender and language differences in Wikipedia, the goal is to scrape a large data set of Wikipedia articles that are likely to be linked and, furthermore, scrape it for multiple languages. Being interested in bias against women in STEM fields, we chose scientists with Wikipedia biographies as our field of study. We are interested in looking at differences between some of the largest languages available in Wikipedia and therefore chose to scrape articles from the English, German, French and Spanish versions of Wikipedia.

Since there is no direct way of scraping the names of all scientists that have a Wikipedia biography, the first part of the data scraping process is to define the names of the scientists we want to include in our analysis. The intuitive first place to look is on Wikipedia's own *List of*-pages that list different topics or people. Ideally, there would be a list of all scientists of all time with a Wikipedia biography; however, such a list does not exist. Instead, we scraped all names from alphabetically ordered lists of notable mathematicians from the WikiProject Mathematics [34] and the sub-lists from Lists of Scientists [35]. All the sub-lists that have been used include scientists from many different fields and nationalities, and they can be found in Appendix A. After all lists have been scraped for names and had their duplicates removed, a list of 28,803 names of scientists remains.

3.1.2 Revisions

Having defined our list of scientists, the next step is to go through the revisions of all articles and scrape and save links to other articles. As mentioned earlier, new revisions of articles can emerge by the minute, every time a sentence is updated or other users negate new updates to articles. This means that saving every single revision of every article is going to be a very time consuming task and not realistic given the length of the list of people that we want to scrape articles from. Instead, we have scraped one revision per year from the beginning of Wikipedia in 2001 to 2024. The goal is not to find every small change in the article content but to catch the changes in links to other scientists.

[1] All implementations and calculations for this chapter can be found in our GitHub repository: https://github.com/ElineBrunke/Gender-and-Language-Differences-in-the-Growth-of-Wikipedia.git

Algorithm 1 Get Revisions

```
function GETREVISIONS(name,scientistnames,year,lang)
    Revisions ← {}
    url ← lang + "wikipedia.org/w/api.php?action=query&list=search&srsearch=" + name +
    "&format=json"
    Content ← OpenURL(url, year)
    Links ← ExstractLinks(Content)
    Revisions ← FilterLinks(Links,scientistnames)
    return Revisions
end function
years ← [2001,2002,...,2024]
languages ← ['en','de','fr','es']
for name in scientistnames do
    for year in years do
        for lang in languages do
            Revisions_df [lang] ← getrevisions(name,scientistnames,year,lang)
        end for
    end for
end for
return Revisions_df [languages]
```

To do so, we have created a function called get_revisions as seen in a simplified form from Algorithm 1. It returns a data frame of one row containing the revisions for that given scientist's Wikipedia page in the given year and language. For instance, get_revisions('Emmy Noether', scientistnames, '2004','en') will return the row [Emmy Noether,2004,[Max Noether, David Hilbert]]. We have chosen to filter the scraped links in the revisions so that they only contain scientists from whom we intend to also scrape data. We are doing this to make sure that we have an accurate representation of the network and that all nodes that are linked to in the network also contain all their out-edges as well.

Now, we created full data frames for each of the 4 languages, looping over years and all names from the list. Afterwards, we merged the one-row data frames returned by the function into one large data frame. After having searched for the 28,803 names in our initial lists, we were able to scrape the content and revisions of 23,843 English articles. For the other 3 languages, the numbers are significantly smaller: 12,281 for German, 7,736 for Spanish and 11,302 for French. We assume that the 4,960 names that are missing from our revision dataset either do not have a Wikipedia page or were misspelled in our list of scientists.

Name	**Year**	**Links**
Emmy Murphy	2023	[Yakov Eliashberg, C. L. E. Moore instructor]
Emmy Murphy	2024	[Yakov Eliashberg, C. L. E. Moore instructor]
Emmy Noether	2002	[]
Emmy Noether	2003	[]
Emmy Noether	2004	[Max Noether, David Hilbert]
Emmy Noether	2005	[Max Noether, David Hilbert]

Table 3.1: Small sample of the English revisions data frame. The data consists of one row for each year a scientist has had a revision. The column Links is a list of the scientists the given scientist linked to that year. This example shows the two last revisions for Emmy Murphy and the first four for Emmy Noether.

A small example of the output is seen from Table 3.1. This is an example of revisions with only a few links, but looking at other people other years, there are instances with long lists in the Links

column. The total length of the revisions data frame is 320,218 rows, but as seen from the example this includes several rows for each person and rows with an empty list in *Links* since the person did not link to anyone from our list that given year.

3.2 Further Data Preparation

3.2.1 Scraping the content of Wikipedia articles

To understand the network we are looking at better, we want to be able to quantitatively investigate topics and fields of all articles and compare these to the network structure. To be able to do so, we have scraped the content of the newest version of the English article of all scientists in our data. This text is initially used to detect birth year and gender, and will later be used to investigate topics and content of the articles.

3.2.2 Detecting gender

To determine gender, we implemented a Naive-Bayse classifier, which is a probabilistic machine-learning model. It is used to classify data into categories using Bayes' theorem to assign articles to classes by calculating the probability of a class given the frequency of specific words in the article [36].
In order to do this, we prepared data using the content of the Wikipedia articles. We counted the instances of the gendered words *she, her, hers, he, him* and *his* as well as *female, women, male* and *man*, and used these as features for the classifier. The classifier is then implemented using the python package sklearn, using its built-in classifier for multinational models. This classifier uses training data where the rows are already assigned to a class to learn the relationship between the features and the classes [37]. We trained and tested it on 600 lines that we manually assigned gender, using an 80/20 train/test split.
After training the model, it was tested, and its confusion matrix was found. It displays the outcome of the test based on the true positives (TP), which in our case is the correctly identified female scientists, the false positives (FP), which is the male scientists wrongly identified as female, false negatives (FN), which is the female scientists wrongly identified as males, and finally the true negatives (TN), which is the correctly identified male scientists.

$$\text{Confusion Matrix} = \begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} = \begin{bmatrix} 27 & 1 \\ 2 & 87 \end{bmatrix} \tag{3.1}$$

The accuracy score for the model was then calculated with this equation

$$Accuracy = \frac{(TP + TN)}{n} = 97.32\% \tag{3.2}$$

which is taking all true female and male predictions and dividing them by the total number of scientists in the test set. The F1 score for the model is calculated with this equation

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \cdot \frac{\frac{\text{TP}}{\text{TP+FP}} \cdot \frac{\text{TP}}{\text{TP+FN}}}{\frac{\text{TP}}{\text{TP+FP}} + \frac{\text{TP}}{\text{TP+FN}}} = 93.02\% \tag{3.3}$$

which is also using precision, which is how many of the samples predicted as positive that were actually positive, and recall, which is how many of the true positives that were correctly identified. Both the accuracy score and F1 score indicate that our model performs excellently.

When deciding on which words to use as features for the classifier, we based the decision both on what we learned from the literature study as well as experimenting. Wagner et al found that there was a language bias towards women in Wikipedia and concluded that the language used to describe women

and men in Wikipedia are remarkably different [4]. When using all words in the articles as features, this difference became very evident, as the test for the classifier revealed that while the accuracy was still large, the precision was very low compared to it. This means that a very large number of women were wrongly classified as men compared to women classified as women. Because of this discovery, we started experimenting with different features, limiting the words that we wanted to use as input and investigating misclassified rows, leading us to only use the specific gendered words as features.

3.2.3 Preparing data and links

To prepare our data to be modelled as a network, we needed to split it into one row per link and detect whether a link was added, removed or remained the same between two timestamps. The first data preparation we had to do was, however, to remove errors in our revisions data frame. Though our initial list of scientists only contained names, we did scrape some articles that were not biographies. This is caused by our search function that searches for the first match on the input. By exterminating the revisions data frame, we did, however, find examples of articles that were, in fact, not about persons. Examples could be 'Economy' or 'Archimedean', but also names of Universities or lists of something. To filter these from our data, we did a lot of manual testing and searching. For instance we printed the 100 names with the most links, where many of these 'fake' biographies appeared. We created two lists to combat this, one called errorlist and one called errorlistcontains. These can both be seen from the Python script **get_dataremodelled** where the two functions used to remodel the data, **prep_data** and **prep_links** are also found.
The first data cleaning we did is done in the function **prep_data**. This function removes all articles where the title is in the errorlist or contains words from errorlistcontains and merges the revision data with the corresponding gender data. This removed 2300 errors from our data. After having used the function on the English revisions our data frame contains 317.918 rows and 16 columns. The columns are as follows: ['Name', 'Year', 'Links', 'he', 'she', 'his', 'him', 'her', 'hers', 'male', 'man', 'female', 'woman', 'Text', 'word_count', 'gender']. The function has naturally also been used on the German, French and Spanish revision data.

The second function created to prepare the data is called **prep_links**. This function takes the full data frame as described above as input. The first alteration done is to add all missing years to the data. This is done since we only have data for the years where the articles have been revised. For instance, the article about the American mathematician Abigail Thompson was created in 2014, and revisions were made in 2015 and 2016. However, in 2017 the article was not edited, which means that Abigail is missing from the revisions data frame in 2017. In 2018, she reappeared as a new edit was done. To fill out the holes in the data, we have created new rows for the missing years containing the links from the latest revision. In the example, the links from the revision in 2016 are duplicated to a new row for 2017.
The second thing we have done to alter the data frame is to explode the links column such that each pair of names and links gets a separate row. Lastly, we created a column **added_or_removed** indicating the difference from last year's revision. This column equals "unchanged", "added," or "removed", depending on the previous year's data.

3.2.4 Detecting birth year

To determine the birth year of all scientists, we use the scraped content from the 2024 versions of the articles. We have implemented a rules-based algorithm, shown in Algorithm 2. The algorithm has been developed in an iterative process, where we, in each step, analysed the detected birth years, found errors and adjusted the algorithm to increase the accuracy. Our base case was searching for the first 4 digit number within the first 1000 characters of the scientist's biography and using this as the birth year.

We then identified two special cases, which is when a scientist was born before or right after the year 0. This is, for example, the case of the ancient Greek astronomers in our dataset. In those cases, the letters BC after a number will indicate if the scientist is born before year 0 and AD or CE after a number will indicate if the scientist is born after year 0, but often less than 1000 years after. The number of digits in the year is, however, not always 4 in these cases. We detected that BC and AD often (but not always) followed a three-digit number being the birth year of the given person, and to find a logic that gives the most accurate numbers in most cases, we decided to go with the three-digit numbers in AD, BC and CE cases.

In our last test, we found that during our text preparations, we removed all special characters, including hyphens. This means articles where the year in which the person lived is denoted as "1921-2002" were changed to "19212002". Therefore, we implemented that the first 4 digits of the first 8 digits were used if there are any and the 4 digit number otherwise. Lastly, we have removed all nan values and values larger than 2024 since this is not a possible birth year for our data.

Algorithm 2 Birth year detection

for scientist ∈ Data **do**
  text ← Scientist(First 1000 characters of the processed Text)
  SEARCH(text, [BC, AD, CE])
  $Year3$ ← SEARCH(text, 3 digit number)
  $Year4$ ← SEARCH(text, 4 digit number)
  $Year8$ ← SEARCH(text, 8 digit number)
  if AD ∈ text **then**
    BirthYear ← $-Year3$
  else if BC ∈ text or CE ∈ text **then**
    BirthYear ← $Year3$
  else if $len(Year8) \geq 1$ **then**
    BirthYear ← $Year8[0:4]$
  else if $len(Year4) \geq 1$ **then**
    BirthYear ← $Year4$
  end if
end for

Firstly, our assumption when using this method is that every scientist is likely to have a birth year in their biography. While this is likely the case for the majority of our dataset, scientists with shorter articles might not have a birth year listed. Scientists might potentially also only have a death year listed or not have confirmed birth or death years, which could be the case for the scientists from before the Renaissance. Furthermore, there is a risk that the first year is not the birth year but a more recent year, for instance, the year the scientist released a famous paper or similar. This does, however, still indicate the time period in which the scientist lived, which is precise enough for the use in this thesis. We expect that few articles have 4 or 8 digits numbers that are not years in the beginning of the text, and in these cases, the algorithm will return false birth years. We do, however, expect this to be very few cases.

Secondly, we assume that the birth year of the scientist will be within the first 1000 characters of the biography. We do this since we are searching the birth in and in most cases, the birth and in some cases death year will appear right after the scientist's full name at the beginning of their biography. We do this instead of searching for 4-digit numbers throughout the entire biography since that might lead to us finding years not related to the scientist's birth or other 4-digit numbers that are not related to years at all.

While we believe that we can use our method without errors in most cases, these assumptions might cause some uncertainty when estimating the birth year of a scientist. After running the algorithm on

the scraped data, we have been able to estimate the birth year of 22.920 scientists. 1048 has been removed since no year was found, and 38 was removed because the assumed birth year was later than 2024.

3.3 The Final Data Set

Our final data consists of 4 tables, one for each language and a small sample is shown in Table 3.2. After all alterations, the English data consists of 626,563 rows and 17 columns. The German data is 298,273 rows, the French 201.807 and the Spanish 129,028.

Name	Year	Link	added_or_removed	Text	Gender	Birth Year
Abigail Thompson	2014	David Gabai	added	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2015	David Gabai	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2016	Martin Scharlemann	added	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2016	David Gabai	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2017	Martin Scharlemann	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2017	David Gabai	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2018	Martin Scharlemann	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2018	David Gabai	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2019	Martin Scharlemann	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990
Abigail Thompson	2019	David Gabai	unchanged	abigail a thompson born 1958 in norwalk connec...	f	1990

Table 3.2: Small sample of the final data set for Abigail Thompson in the years 2014-2019. The data consists of one row per link between two persons from the scientist list. Note that the birth year is an estimate and in the case of Abigail not precise but an indication of living period

The data is now prepared to create a network of all scientists in the four different languages, as well as analyse the article's content further.

3.3.1 Bias in lists of scientists

As mentioned earlier, the first task in the data scraping process was to define the scientists in our network. In doing so, we found some unexpected biases in the lists of scientists we scraped to get the list of all scientists' names for our data set. When going through the list of lists of scientists, we found both lists of persons of a given science and a list of women of that same science. For instance, there is both a list of List of Chemists and a List of Women in Chemistry. The list of women in chemistry is, however, not included in the list containing the lists of scientists.

Otherwise, this could be a fine structure if there was a significant overlap between the two lists and the female chemists were simply just included in both lists. This is, however, not the case since only 26 of the 140 female chemists are included in the list of chemists. By using our own gender function to detect the gender of the members of both lists, we found that the list of women in science consists of 134 women and 5 men and the last person we were not able to find was the Wikipedia article. For the list of chemists, we found that 534 were men and 136 were women. When looking at the list of chemists excluding the names from Women in Chemistry, there are 110 women and 533 (according to the gender found by our own gender classifier), indicating that not only are female chemists missing from the list

of chemists, but some are also missing from the list of women in chemistry.

This both indicates a bias towards women in the way Wikipedia defines who are chemists which makes it difficult to find a comprehensive overview of women in science, and it creates uncertainties in the data used to investigate other biases in Wikipedia - including this thesis.

3.3.2 Gender and counts of gender-specific words

A small sample of the output of the gender classifier is seen in Table 3.3. The classifier categorises 19,616 of the scientists as male and 4,237 as female, corresponding to 82.24% males.

Name	he	she	his	him	her	hers	male	man	female	woman	Text	Word Count	Gender
Ignaz Schütz	1	0	0	0	0	0	0	0	0	0	ignaz robert schtz 1867 bezov moravia 1927 br...	111	m
Thomas-François Dalibard	3	0	1	0	0	0	0	0	0	0	thomasfranois dalibard french pronunciation tm...	256	m
Charles H. Zeanah	2	0	1	0	0	0	0	0	0	0	charles h zeanah jr is a child and adolescent ...	557	m
Hideki Imai	5	0	2	0	0	0	0	0	0	0	hideki imai imai hideki born 1943 in shimane...	195	m
Anna Romanowska	0	2	0	0	2	0	0	0	0	0	anna b romanowska is a polish mathematician sp...	131	f
Immanuel Bonfils	3	0	1	0	0	0	0	0	0	0	immanuel ben jacob bonfils c 1300 1377 was a ...	489	m
Myhailo Yadrenko	14	0	14	0	0	0	0	0	0	0	myhailo yosypovych yadrenko ukrainian was b...	1196	m
Samuil Shatunovsky	12	0	8	0	0	0	0	0	0	0	samuil osipovich shatunovsky russian 25 mar...	306	m
Johan Frederik Steffensen	3	0	1	0	0	0	0	0	0	0	johan frederik steffensen 28 february 1873 in ...	183	m
Tu Youyou	0	16	1	0	17	0	0	0	2	0	tu youyou chinese pinyin t yuyu born 30 decem...	1544	f

Table 3.3: Small sample of the results from the gender classifier. The columns 'he', 'she', 'his', etc., are the number of times each word appears in the text string. The word count is the total number of words in the article and Gender is the classified gender

In the process of detecting the gender of all articles, we found that identifying the women was more difficult than identifying the men, leading us to look more at the counts of gender pronouns. The average percentage of gendered words in the articles is seen in Table 3.4.

Percentage female words in female articles	93.15 %
Percentage male words in female articles	6.85 %
Percentage female words in male articles	1.91 %
Percentage male words in male articles	98.09 %

Table 3.4: Average percentage of female and male words in female and male articles. The percentages are calculated as the number of female words ("she","woman" etc.) divided by the total number of gendered words for female articles and the opposite for male articles. An article classified as a woman with 4 female words and 1 male will result in a percentage of $\frac{4}{5} \cdot 100 = 80\%$

By counting the number of female and male pronouns, we see that on average, female articles have more male pronouns than male articles have female pronouns. To investigate if this difference is significant, we have used a Two-sample Confidence Interval for the difference between two means by non-parametric bootstrap [38]. We do this by formulating a null hypothesis and attempting to reject it by resampling the dataset, calculating the differences in the sampled means and then analysing the distribution of them. The null hypothesis we want to test is that the mean of female words in female articles and male words in male articles is not the same, which can be described as the difference between them being zero.

$$H_0 : \mu_{\text{Male words in male articles}} - \mu_{\text{Female words in female articles}} = 0 \tag{3.4}$$

The null hypothesis is seen from Equation 3.4, and being able to reject this would mean that the two means are indeed different from each other and that the difference is statistically significant.

We are performing this test using sampling since the group of male articles is so much larger than the group of female articles. We are defining 1000 sets of randomly selected male and female articles and their word percentages. Each of these sets contains 1000 male and 1000 female articles, of which we find the mean of the difference between the percentage of male words in male articles and female words in female articles for each set. Finally we can compute the confidence intervals using the set differences and use those to conclude if we should reject the null hypothesis. The confidence intervals that we computed can be seen in Figure 3.1 along with the differences that we calculated for all 1000 sets.

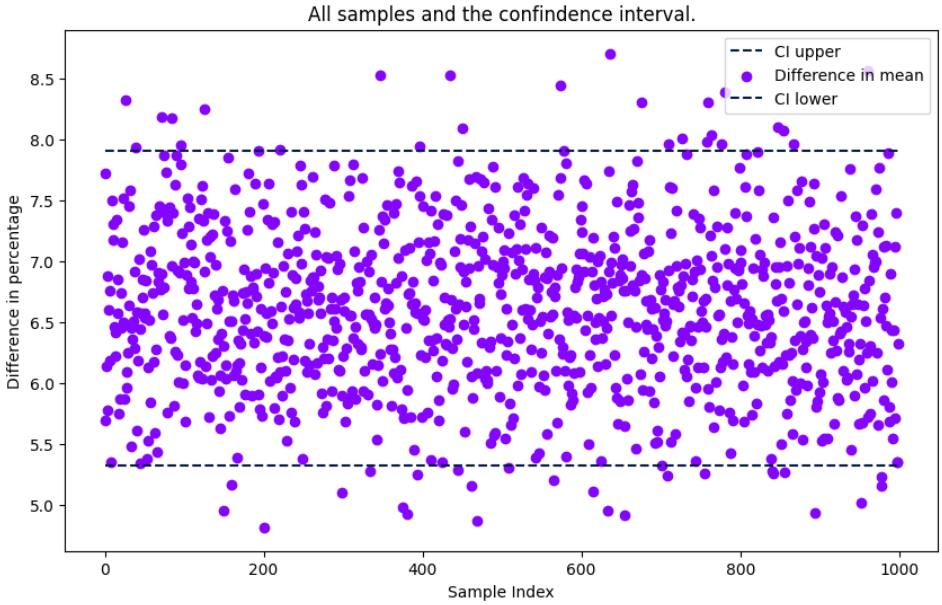


Figure 3.1: The 1000 differences in means, as well as their 2.5 and 97.5 confidence intervals. Each difference is found by taking a sample from the male and female populations respectively and finding the difference between the means of the share of gendered words that corresponds to the gender of the person the article is about.

For the two means to be the same, we would expect the confidence interval to centre around zero. This is, however, not the case, since the lower limit of the confidence interval is 5.39 and the upper limit is 7.86. From this, we can conclude that the null hypothesis can be rejected and that the difference between female words in female articles and male words in male articles is statistically significant.

In addition to the count of gender pronouns a word count of each article has also been done for the English version of the articles. Across all articles in our data, there is an average word count of 820.39 words per article. Dividing the data into female and male articles, the male articles have an average word count of 872.22 words, and the female articles have an average word count of 621.17 words.

3.3.3 Birth Year

The mean birth year of our network members is 1874 and the birth years range from -753 (753 BC) to 2019. The lower quartile is 1863, and the upper is 1949, meaning that 50% of the persons in our network were born between these two years. From Figure 3.2, box plots of the gender distribution are seen. In general, we can conclude that the women in the network are born later than the men, and in

a smaller range, for the women, 50% of the scientists are born between 1913 and 1980 while 50% of the men are born between 1852 and 1944. The average birth year of the women in our network is 1925, and 1850 for men.

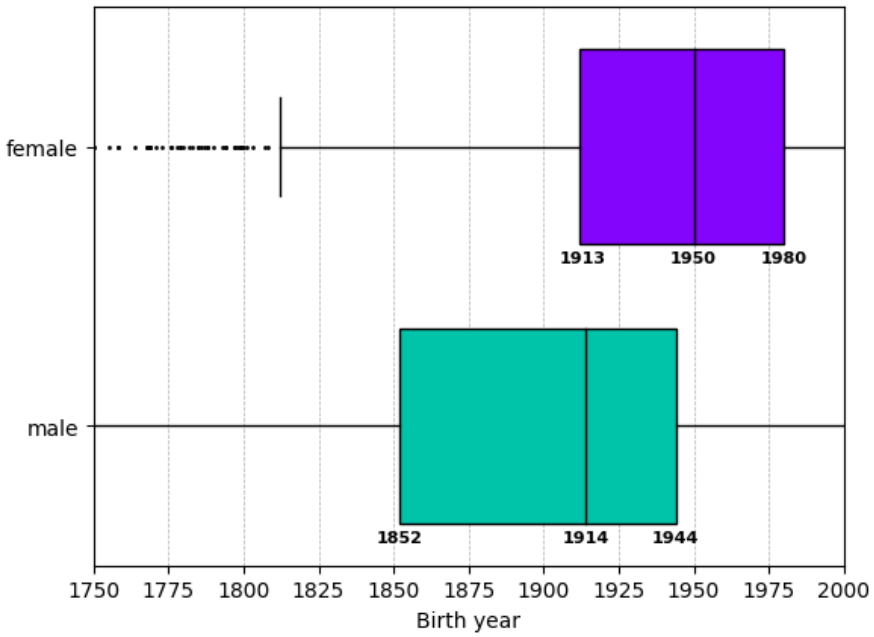


Figure 3.2: Estimated birth year of all persons in the network split by gender. The median birth year for women is 1950, while it is 1914 for men. Despite being lower, the estimated birth years are also wider spread for the men

CHAPTER 4

Network Construction and Analysis

This chapter covers the construction of a network based on the data scraped in the previous chapter. We will both cover the methods and theory used to construct, visualize and analyse the networks as well as the findings we have discovered in our analysis of the English network. The network is analysed by looking at the structure of the network, the content of the articles and lastly, the growth over time.[1]

4.1 Understanding Networks

Throughout the analysis of our thesis, we will model the scraped Wikipedia data as a network. A network is a collection of nodes and the edges between them. They can be either undirected, where if node a is connected to node b then node b is also connected to node a, or directed, where the direction of the connections between nodes matters and a connection from a to b does not ensure that there is a connection from b to a [39]. When representing Wikipedia articles as a network, it will, in most cases, be the obvious choice to use a directed graph as articles link to other articles, but these articles will not necessarily link back.

The network can be considered homogeneous or heterogeneous, where the homogeneous network is made up of the same type of nodes and links, and links can occur between any pair of nodes. On the other hand, a heterogeneous network can have different types of nodes and links, and different rules as to which types of nodes can connect [40].
Given these circumstances, a Wikipedia network would, aside from being directed, also be considered mostly homogeneous since there are no different types of nodes or edges. However, there are still some heterogeneous aspects to this network, because of Wikipedia's rules in regards to editing. Wikipedia articles are written about a variety of different topics and fields, and links can, therefore, not freely occur between any random pair of nodes in the network.

4.2 Network Construction and Visualisation

For the graph construction, we are using the Python package NetworkX along with its online documentation [41]. Using this package allows us to establish a graph and then add nodes and edges using the revisions from our data frame. When creating directed graphs, we will start by creating a NetworkX graph and then iterate over all rows of the data frame. From the data frame, we will use the 'Name' and 'Link' as the source and target. If the source has not yet been added to the graph, we will add it, and the same goes for the target. Other than the name of the source or target, their gender and year are also added to the node when it is added to the graph. After ensuring that both the source and the target are in the graph, an edge is added between them.

[1] All implementations and calculations for this chapter can be found in our GitHub repository: https://github.com/ElineBrunke/Gender-and-Language-Differences-in-the-Growth-of-Wikipedia.git. Here html-files for the Sankey plots as well as selected times series plots are found

Algorithm 3 Graph Creation

```
G ← ∅
for row ∈ dataframe do
    if source ∉ G then
        G.add_node(source,source_gender,year)
    end if
    if target ∉ G then
        G.add_node(target,target_gender,year)
    end if
    G.add_edge(source,target)
end for
```

The method for creating graphs is shown in Algorithm 3, and this will be used every time we create a graph throughout our thesis. It will be used with data frames that have been filtered on year so that the created graphs will be yearly snapshots. When investigating growth over time, we will, therefore, create separate graphs for each of the years and then compare the nodes and edges using the NetworkX node and edge attributes.

The graphs created in this thesis are large and visualizations can be difficult to comprehend. Therefore the graph visualization tool Gephi has been used. The previously mentioned NetworkX package has a function that can convert a graph to a Gephi-file that can be uploaded to Gephi. In Gephi, the graphs have been visualized using the build-in function OpenOrd. OpenOrd is good for representing large graphs. It does, however, expect undirected, weighted graphs [42]. Our graphs are the opposite, directed and unweighted so OpenOrd was not chosen as the obvious choice for our data, but because it gave the best results in the shortest runtime.

4.3 Network Construction and Visualisation Results

The 2024 network consists of a total of 15,690 nodes and 45,745 edges, meaning that each node, on average, has 2.92 in-going edges and 2.92 out-going edges. This also means that out of the 23,843 articles we managed to scrape with the get_revisions function, 8,153 people are not in our graph due to the person not linking to other people from our list. The Network has been illustrated using Gephi and can be seen in Figure 4.1. By illustrating the graph, it is clear that the network mainly consists of male nodes and that the largest nodes (with the highest in-degree meaning that many other articles refer to the given article) are men.

The development in the number of edges and nodes over time is seen in Figure 4.2. From this, we see the largest growth in both edges and nodes in 2007, 6 years after the foundation of Wikipedia. From 2007 and 10 years forward the yearly growth decreases, but in 2018-2020 there was a small peak again. After 2020, the growth has been falling for both the number of nodes and edges.

Figure 4.1: The English Network coloured by gender. Node size is based on the in-degree of the node and the edge colour is coloured as the target node. Green nodes are men, purple nodes are women. We see that not only are very few nodes female articles, but the female nodes are also small, meaning that they have few ingoing edges.

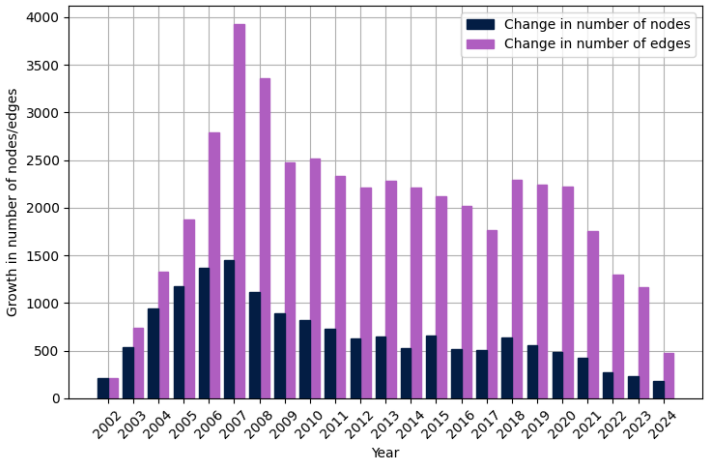


Figure 4.2: The number of edges and nodes added to the graph each year from 2002 to 2024. Each bar represents the difference in the number of edges or nodes compared to the year before. If 1500 new nodes have been added but 2 removed this will correspond to a total change of 1498 that year. We see a peak in the number of nodes and edges added in 2007, where the network has almost 4000 edges and almost 1500 nodes more than it did in 2006.

The connectivity over time measured as the number of edges per node is seen from Figure 4.3. This figure follows the trends we saw in Figure 4.2: in the first years of Wikipedia the number of edges

per node was small, but with the growth of the Wikipedia network more edges were added and over time the network does not only grow in regards to the number of nodes but also gets better connected. However, the development slowly stagnates from 2020 and onwards, and between 2023 and 2024 there is even a small fall, meaning that more nodes than edges are added to the network that year.

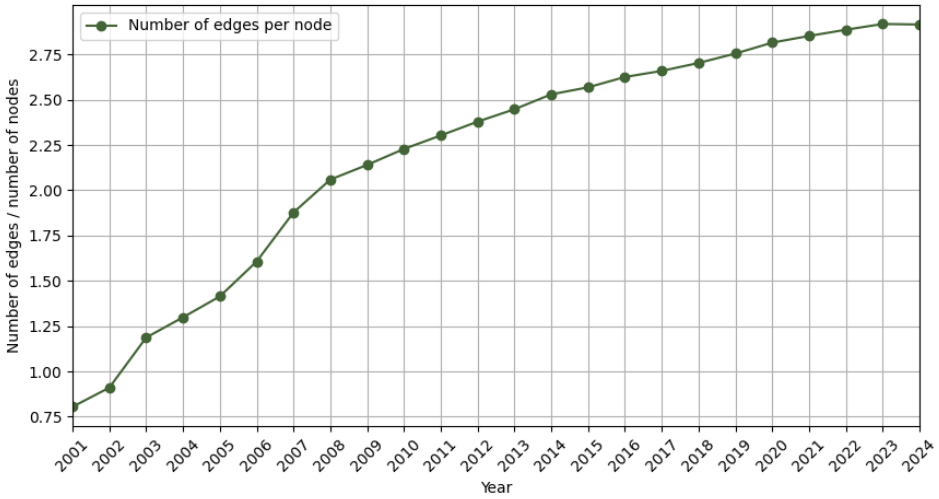


Figure 4.3: Connectivity measured as the number of edges/the number of nodes. In the first approximately 10 years, the number of edges per node grows rapidly, meaning that the network becomes more and more connected. This pattern continues until 2023 but at a slower pace.

4.4 Network Structure

4.4.1 Degree and degree distribution

An important measure when analysing networks is the degree. The degree is simply how many other nodes the node is connected to [39]. In this thesis, we use the terms degree, in-degree and out-degree. In the context of our Wikipedia network, the in-degree of a node is the number of ingoing edges a node has, the out-degree is the number of outgoing edges, and the degree is the total number of edges. The reason we use in-degrees when visualising the graph is that the number of other articles that link to a given person reveals more about the importance of the person than how many articles the person links to. The degree distribution represents the probability p_k that a random node in the network has degree k. The degree distribution is defined in the Network Science Book [39] Chapter 1 as

$$p_k = \frac{N_k}{N} \tag{4.1}$$

where N_k is the number of nodes with degree k and N is the total number of nodes. The degree distribution is often interpreted by plotting it in a log-log plot. This way it is easy to determine if the distribution follows a power law distribution. Networks that do follow a power law distribution are called scale-free networks, and networks such as the World Wide Web, social networks and also networks of Wikipedia articles are most often characterized as such. We do, therefore, also expect our network to be categorized as a scale-free network.

4.4.2 Robustness and k-core

If our network exhibits scale-free patterns, we also have some expectations about its robustness. Scale-free networks are categorised by highly connected hubs, and new nodes in the graph are more likely to

connect to existing nodes with higher connectivity. Because of this, the scale-free network would be robust against random failure, since it would be statistically more likely that a low-degree perimeter node would be removed since they make up the majority of the network. While we would expect that the network would still persist if low-degree nodes were targeted and removed, the highly connected nodes could be an Achilles' Heel if they were attacked and removed [39].

To get an idea about how connected our network is, we will use a k-core algorithm. This is used to identify what parts of the network are more resistant to being removed from the graph when one or more of their neighbours are removed and, subsequently, which of our nodes are located in our network's perimeter. The k-core of a graph refers to a sub-graph of a graph where all nodes have the degree of at least k. We can use it to identify how many nodes impact the overall connectivity more, as well as the level of robustness in the network [43].
We will find the k-cores for the graph of our final network from 2024. Since our graph is directed, we have chosen to implement a k-core algorithm that considers both the in and out degrees of the nodes. The algorithm works by finding the degree of each of the nodes and removing those that have a degree smaller than k. This process is repeated until the nodes are no longer removed from the graph. In order to analyse the robustness of the graph, we will also keep track of the ratio of female and male nodes that are removed when creating the k-core of the graph, as well as the number of nodes in the k-core compared to the initial number of nodes.

4.4.3 Communities

In order to investigate the connections and relationships in our network, we can identify which nodes have a higher likelihood of connecting to each other than to other nodes. The nodes that are highly likely to connect to each other are grouped in the same community. After computing the communities, we can use them to get an overview of the structure of the network and natural grouping, identifying interesting sections and, over time, see if subsections merge together or grow apart.

When choosing a method for community detection to use with our data, we have considered a few things. Firstly, we would like an algorithm that can handle our network type, which is a big network of scientists that we suspect are working together within and across fields. We want to be able to identify how they are working together and the community detection method should capture exactly that underlying structure of the network.
Secondly, because our dataset ends up containing more than 23,000 nodes, we prefer a fast algorithm since we want to create communities for all our graphs over the years for all languages. Another benefit of using a fast method is that we were able to make changes in the data and rerun the community detection in case of errors in our data-preparing processes.

Therefore we have chosen to use InfoMap, which identifies communities by using the concept of a random walker moving through the network. It seeks to describe the path of the walker as efficiently as possible using the fewest number of symbols by taking advantage of the fact the random walker will tend to get "trapped" and stay longer in densely connected parts of the graph, highlighting the communities. It does this by minimizing the map equation

$$L = qH(Q) + \sum_{c=1}^{n_c} p_c \cdot H(P_c) \tag{4.2}$$

where first term $qH(Q)$ provides the necessary bits for the walker moving between communities and the second term $\sum_{c=1}^{n_c} p_c \cdot H(P_c)$ provides the necessary bits for the walker moving within communities. By minimizing the equation L, InfoMap finds the shortest description of the random walk and, by this, the community structure of the network [39].

When creating the communities we are using the InfoMap Python Package, which has a model that initialises an InfoMap framework and lets us populate it with the nodes and edges from our networks [44]. When our networks have been inserted, we can run InfoMap to define the communities and create a community summary that contains the communities and their members, along with statistics about their size and percentage of women.

While we have chosen this method based on the above rationale, we also recognise that it lacks in certain aspects. An example of this is overlapping communities, which InfoMap cannot detect. When taking the type of network we are working with into account, we do however expect that some of our scientists could be members of several communities, for instance, a mathematician could have discovered some things that could be used by both physicists and chemists and therefore share many connections with these communities, while still being relevant for a community of mathematicians.

4.4.4 Strength of communities

A way to evaluate how connected the communities are is to determine whether they are strongly or weakly connected. In the Network Science Book [39] a strong community is defined as a community where each node in the community is more connected to nodes in that given community than to nodes from the remaining network. Mathematically this is defined as

$$K_i^{int}(C) > K_i^{ext}(C) \forall i \in C \tag{4.3}$$

where $K_i^{int}(C)$ is the number of neighbours node i has within community C and $K_i^{ext}(C)$ is the number of neighbours node i has that are not in community C. Neighbors are defined as nodes node i share an edge with, regardless of the direction of this edge.
This is a rather restrictive definition and can be relaxed by using the definition of weak communities. A weak community is defined as a community that meets the following definition:

$$\sum_{i \in C} K_i^{int}(C) > \sum_{i \in C} K_i^{ext}(C) \tag{4.4}$$

This method sums the number of edges from community nodes to other nodes in the same community and to other nodes in the network. With this definition, a community where some of the nodes are less connected in their own community than with the rest of the graph can still meet the constraints as long as the remaining nodes have a corresponding amount of edges inside the community than with the rest of the graph.

Based on these two definitions from the book, we have in this thesis introduced our own strength measure to be able to determine how far from being strong a community is. The measure is seen from Equation 4.5 and is the fraction of members in a community that connect more to nodes in their own community than nodes from the remaining network.

$$Strength(C) = \frac{n_s(C)}{n(C)} \tag{4.5}$$

where $n_s(C)$ is the number of nodes in the community with more internal edges than external edges and $n(C)$ is the total number of nodes in the community.

In other words, strength measures the proportion of nodes that connect more to/from nodes within their own community than the rest of the network.

4.4.5 Text Processing

To find out what each of the detected communities has in common we have merged all text from the Wikipedia articles in each community to one string per community. To be able to count the frequency

of all words in all articles from a given community, the words have first been converted to lower case and stemmed. Stemming means that each word in the text has been reduced to its root form. An example of how this works is the words *scientific* and *scientifically* that both are reduced to *scientif* when stemmed. Besides from stemming the text all names of scientists in the data set have also been removed from the text. This is done as we already know the names of the scientists in each community and they are therefore not giving any new information about the given community. Often, one would remove stop words such as *"that"*, *"or"* and so on, but in this case, we have decided to keep those words in our text. Our main purpose with the text besides the gender and birth year detection we did during our data preparation, is to compute TF-IDF scores. TF-IDF already takes into account how often a word occurs in other texts, and if the analysis returns a stop word we assume that this given stop word is significant for the text in question.

4.4.6 TF-IDF

To gain information about what words are significant for the articles and to get a better idea about the topics and fields of the communities, we have calculated the TF-IDF scores for all words in each community. TF-IDF score is a combination of the term frequency (TF), how often each word appears in the community text and inverse document frequency (IDF), the relevance of each word measured by how often the word occurs in other communities. The term frequency is calculated for all terms in each of the communities with the following equation

$$TF_{t,c} = \frac{f_{t,c}}{\sum_{t_s \in c} f_{t_s,c}} \tag{4.6}$$

where the numerator is the count of the term t in the community c and the denominator is the sum of the counts of all the terms t_s in the community c. The inverse document frequency is calculated with the following equation

$$IDF_t = \log(\frac{\mathbf{C}}{n_{t,c_s} + 1}) \tag{4.7}$$

where $\mathbf{C}$ is the total number of communities and n_{t,c_s} is the number of communities, that contains the term t. The final TF-IDF score of the term t in community c is calculated as seen below

$$TF\text{-}IDF_{t,c} = TF_{t,c} \cdot IDF_t \tag{4.8}$$

Having calculated the TF-IDF-scores for all words and communities, the most significant words for each community are those with the highest score [45].

4.4.7 Movement between communities

To visualise movement between communities over time and between language versions, we use Sankey diagrams. Sankey is a visualisation tool where the flow between groups is shown [46]. In our case, it is used to show the movement of persons between communities across time and language versions. To create these plots we have used *graph_objects* from Plotly. The code used for our Sankey graphs is inspired by an example from plotly.com [47].

To visualise the similarities between communities across languages, we will also use Confusion matrices. These are visualisations that typically are used to detect the performance of classifiers, as seen with our gender classifier in Equation 3.1. In the context of comparing communities, it is, however, used on a large scale as a $n \times n$ matrix with the number of common elements as values. These values are coloured after size, making it easy to interpret and localise the maximum values [48].

4.5 Network Structure Results

4.5.1 Degree distribution

The degree distribution for male and female nodes is seen from Figure 4.4. We can conclude that the female nodes are poorly represented among the nodes with the largest in and out degrees. The difference is most pronounced when looking at the in-degree indicating that the female nodes are less likely to be linked to than they are likely to link to others. This is also evident when looking at the average degrees where the average in-degree is 1.30 for women and 3.20 for men. The difference in out-degree is slightly smaller, 1.97 for women and 3.08 for men.

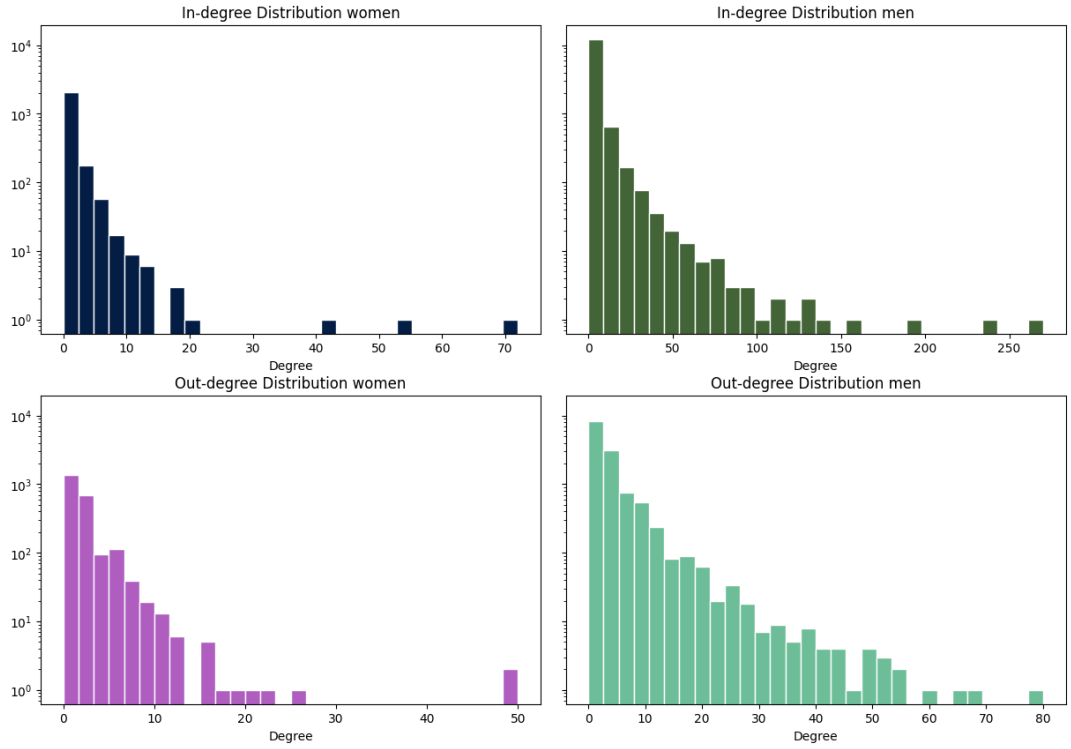


Figure 4.4: In- and out-degree distributions for female and male nodes of the 2024 network. The network has a few very connected nodes with in-degrees larger than 100 and mostly consists of nodes with an in-degree smaller than 50. Very few of the female nodes have a in-degree larger than 50 while the plot shows a longer tail for the male nodes. For the out-degree, the two distributions are more alike, but the male nodes are still, in general, better connected.

From Figure 4.5, two log-log plots of the degree distribution are seen, one for the in-degree distribution and one for the out-degree distribution. We see from these plots that the degree distribution follows a power law distribution, and we can, therefore, conclude that the network, as expected, can be classified as a scale-free network.

By looking at the degree distribution, we can conclude that the network consists of a few very connected nodes with an in-degree larger than 100 and many nodes with an in-degree of 1 or less. This difference in degree will be investigated further in the next section.

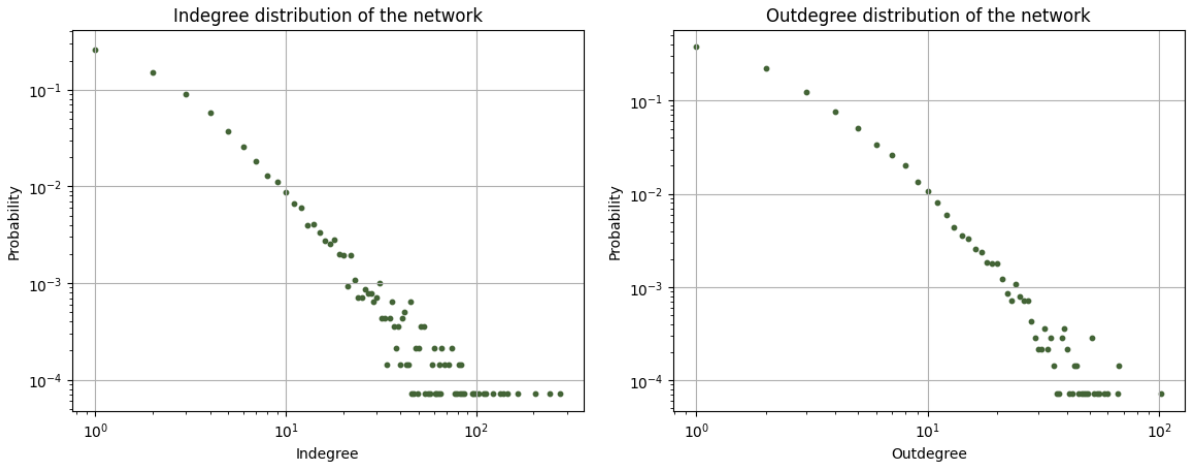


Figure 4.5: Log-log plot of degree distribution for both in- and out-degree. Both degree distributions follow a power law distribution, and we can therefore conclude that the network is scale-free.

4.5.2 Robustness and k-core

To analyse the structure and connectedness of the network further, we computed both a 1, 2 and 3-core for our network and counted the percentage of the nodes removed, as seen in Table 4.1.

k core	Nodes left	Male nodes removed	Female nodes removed	Female nodes left
1	74.6%	23.7%	35.1%	13.1%
2	50.2%	46.6%	67.8%	9.35%
3	34.6%	62.5%	81.9%	7.87%
4	22.0%	75.8%	90.7%	6.36%

Table 4.1: The stats calculated after finding the 4 k cores of the graph. The nodes left are the number of nodes in each of the k cores compared to the number of nodes in the original graph. The male and female nodes removed are the percentage of the male and female set of nodes that are removed in each of the k cores. The female nodes left are how much female nodes make up of the k core.

The ratio of the nodes removed increases rapidly throughout the four k cores of the graph. This indicates that the graph is not robust to the removal of the lower degree nodes. This aligns with the previous results, where we found that the degree distribution of the graph is highly skewed, with a few high-degree nodes and many low-degree nodes.
Female nodes are removed at a faster rate than the male nodes, only making up 6.36% of the final k-core of the graph. In each k-core, the female population gets removed until only 9.3% of the female nodes are left in the 4-core of the graph. This is a very small fraction left compared to the fraction of male nodes left, which in the 4-core is 24.2%. If the male and female nodes were equally connected in the graph, we would expect them to get removed at the same rate throughout the k-cores, this is not the case. This also aligns with what we found when comparing the male and female nodes' degree distribution, where the degree distribution for the female nodes was much slimmer. It, therefore, also seems like the robustness of the graph depends heavily on the male nodes and way more than on the female nodes, as they remain in k cores for much longer than the female nodes.

4.5.3 Community detection

The networks of all 24 years are grouped into communities using InfoMap. For the 2024 data, this results in 526 communities with a mean community size of 29.83 members. 512 of the communities has less than

50 members and will not be considered in the further analysis. The size, gender distribution, average in- and out-degree and strength of the remaining 14 communities are seen in Table 4.2. From these statistics, we can see that the communities differ in size, proportion of female members and connectivity measures. The largest community is Community 2, and this community is among the most female-dense communities. Looking at the average in- and out-degree, it is difficult to spot the differences with the naked eye. However, the in-degree of Community 1 is slightly higher than for the remaining communities, indicating that members of this community are linked to more often than non-members. Looking at the strength measure defined in Equation 4.5, the largest communities (1,2,3,4) seem to be more connected within their own communities than some of the smaller communities (16,9,10,11).

Community	**Size**	**Female members**	**Average in-degree**	**Average out-degree**	**Strength**
1	2517	12.59%	4.32	3.85	84.96%
2	2820	20.00%	3.41	3.28	84.65%
3	2097	8.63%	3.20	3.19	87.25%
4	1528	6.68%	3.38	3.27	90.02%
5	1302	16.74%	3.47	3.38	72.72%
6	1292	12.69%	3.31	3.23	80.61%
7	1015	17.73%	3.58	3.42	78.34%
8	710	20.70%	3.30	3.11	78.07%
9	175	9.71%	3.38	3.19	60.88%
10	280	16.43%	3.27	3.09	75.43%
11	128	15.63%	3.63	3.27	65.05%
12	109	13.76%	3.20	3.16	55.70%
13	157	15.29%	3.16	3.02	82.68%
16	58	12.07%	3.47	3.02	77.33%

Table 4.2: Size, percentage of female members, strength, average in- and out degree of all communities larger than 50 members. The percentage of female members differs from community to community, with communities 2 and 8 as the communities with the largest percentage of female members and communities 3 and 4 having the lowest percentage.

Based on these preliminary statistics, we have selected 5 focus communities, communities 1,2,3,4 and 8. These communities are chosen based on their size and gender distribution as well as the connectivity measures, with a goal of analysing the communities with the largest and smallest percentages of female members.

From Figure 4.6 the graph coloured after the focus communities is seen. Other communities than the 5 focus communities are grey. We see that the communities created using InfoMap also tend to be grouped in the Gephi graph. We also see that even though community 2 is the largest community we found, it does not contain as highly connected nodes as the other communities, even though its average in- and out-degree does not differ considerably from the other communities. The graph is however very large, and to gain more information about the identified communities, we will analyse these further.

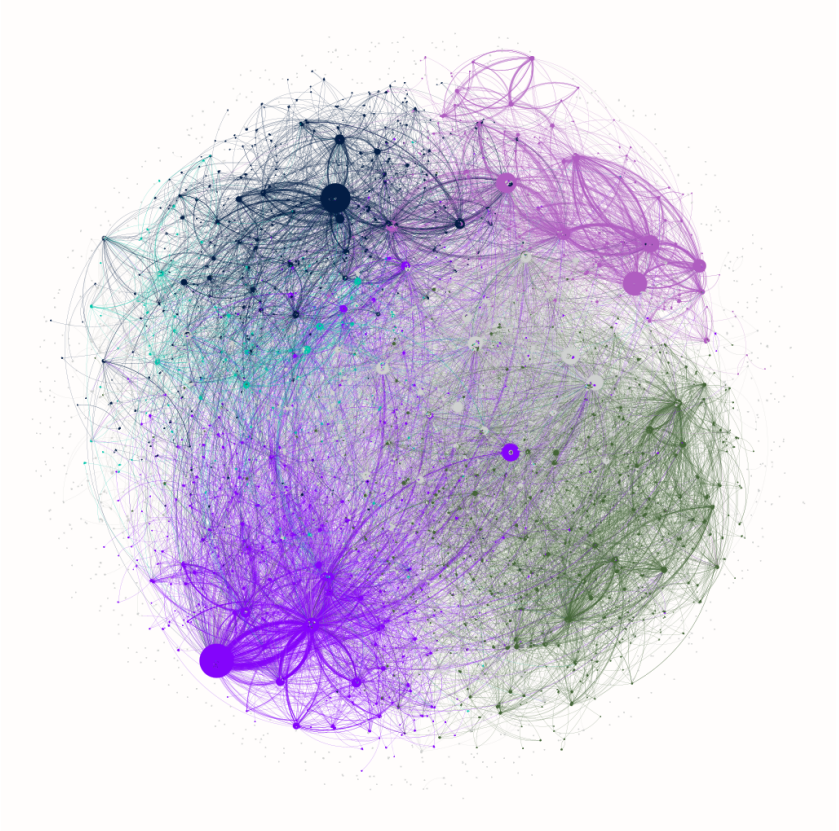


Figure 4.6: The English network coloured after focus communities with community 1 being violet, community 2 being dark green, community 3 being dark blue, community 4 being light purple and community 8 being light blue. Edges are coloured after their target node and the node size is defined by in-degree.

The percentage of members of each focus community that has been added to the graph over time is seen in Figure 4.7. This is based on the year that the articles in the community in 2024 were first added to Wikipedia. From this plot we see that the members of community 2 are added later to Wikipedia than the members of communities 3 and 4, meaning that the articles in 3 and 4 in general are older than articles from other communities. These two communities are also the communities with the lowest birth years, so there seems to be a connection between birth year and when an article is written, which corresponds with what one could expect.

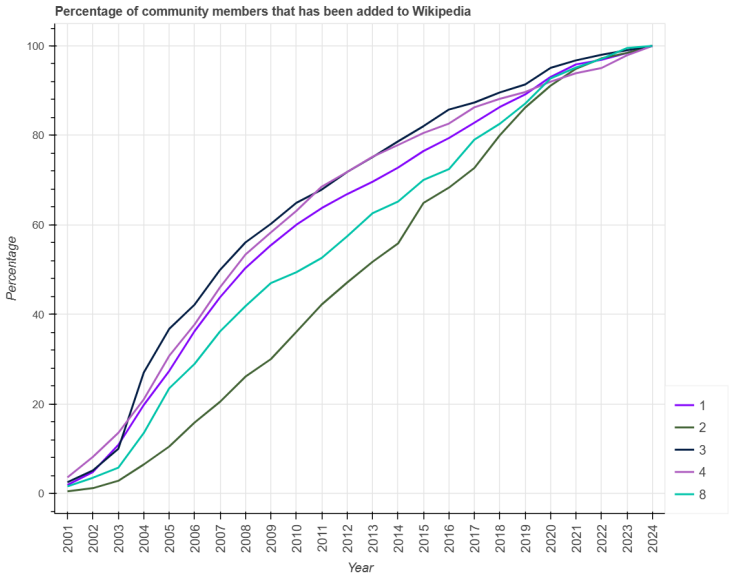


Figure 4.7: The percentage of the focus community members that has been added to the graph over time. The members are added at a very different rate from community to community where the members of community 2 are added later than the members of community 3.

Looking at the 5 focus communities the fraction of female members is very different from community to community. As seen in Figure 4.8, all five focus communities have less than 10% female members in the first 12 years of Wikipedia. After 2013, the fraction of female members, however, grew very differently for the 5 communities. Where communities 2 and 8 grow rapidly in the following years, 3 and 4 stagnate, and 1 grows slowly. The end result is 8 and 2 having above 20% and 4 and 3 still being below 10%.

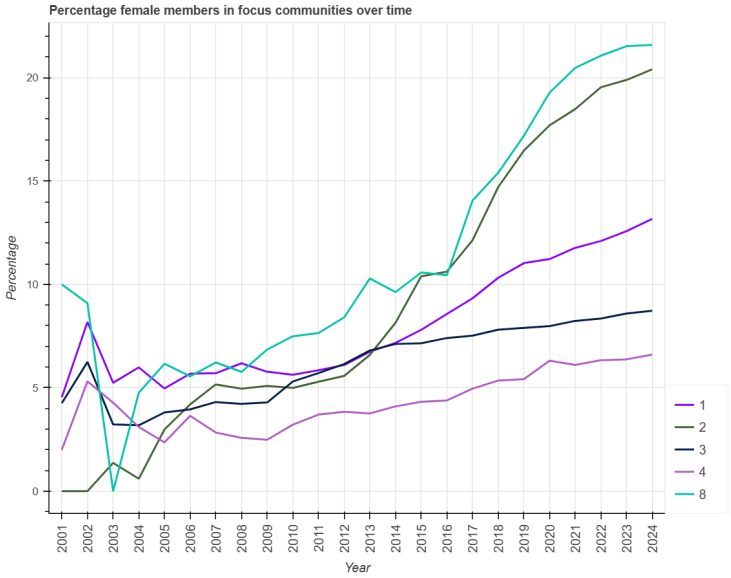


Figure 4.8: Development in gender distribution for the 6 focus communities. The percentage of female members is low for all communities until 2016 where especially communities 2 and 8 were growing fast towards the 20%

Looking further into the time period in which the members of each community lived, the distribution of estimated birth years is seen in Figure 4.9. As seen from the average and median birth years, Community 4 is the most remarkable in this context. Not only is the median significantly lower than for other communities, but the estimated birth years are also, in general, wider spread from the 25th percentile 1280 to the 75th percentile 1759, a gap of more than 400 years. The community with the smallest gap is surprisingly the largest community, Community 2, with a 25th percentile of 1927 and 75th percentile of 1966, only 39 years between.

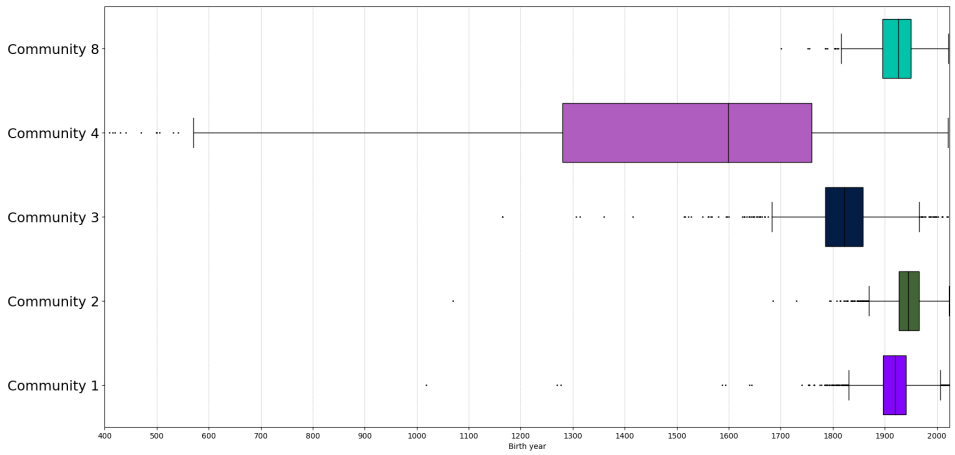


Figure 4.9: Estimated birth year of persons in focus communities. Community 4 is the most significant with older members than the other communities. 25th percentile, median and 75th percentile of all communities with 50 members or more is seen from Appendix B

These differences in when the members of each community were born and, thereby, the time period in which they lived indicate that the community member has something in common, and it is therefore highly relevant to look further into how the differences between communities can be described.

4.5.4 Most connected nodes

The top 10 most connected nodes in each of our focus communities are measured by in-degree, and the results are seen from Table 4.3.

Community 1	**Community 2**	**Community 3**	**Community 4**	**Community 8**
Albert Einstein	Richard Courant	Charles Darwin	Aristotle	Richard Dawkins
John von Neumann	Alexander Grothendieck	Alexander von Humboldt	Isaac Newton	Karl Pearson
Niels Bohr	André Weil	Carl Linnaeus	Euclid	Ronald Fisher
Enrico Fermi	Shing-Tung Yau	Georges Cuvier	Ptolemy	Stephen Jay Gould
Werner Heisenberg	Andrey Kolmogorov	Louis Agassiz	Galileo Galilei	Ernst Mayr
Arnold Sommerfeld	Barry Mazur	Charles Lyell	Leonhard Euler	Jerzy Neyman
Max Born	Edward Witten	Joseph Banks	Archimedes	Daniel Dennett
Richard Feynman	Solomon Lefschetz	Rudolf Virchow	Johannes Kepler	Gregor Mendel
Linus Pauling	Michael Atiyah	Michael Faraday	René Descartes	Julian Huxley
Max Planck	Élie Cartan	Alfred Russel Wallace	Avicenna	Steven Pinker

Table 4.3: Top 10 most connected nodes in each focus community measured by in-degree in the 2024 graph. Note that the top 10 most connected in all focus communities are male, including the communities with the largest percentage of female members

We see that despite some communities having more than 20% female members there is no women among the top 10 most connected to nodes in any of the focus communities. Several well-known

scientists are seen in the top 10 lists, for instance, Albert Einstein and Niels Bohr in community 1 and Charles Darwin in community 3. Looking at who is in each community gives an indication of what the members have in common, but to gain more insights into this, we will take a look at the content of all articles in the communities.

4.5.5 Community content

The top 30 words found by the TF-IDF calculations for each of the focus communities are seen in Table 4.4. In Appendix C the top words for the remaining communities can be found.

Community ID	Top 30 words by TF-IDF score
1	quantum, nuclear, atom, that, hi, particl, she, energi, laboratori, had her, it, theoret, physicist, physic, would, be, thi, space, mechan institut, electron, develop, chemic, soviet, award, scientif, war, cambridg, experi
2	mathemat, algebra, geometri, topolog, differenti, conjectur, she, manifold, displaystyl equat, mr, space, function, math, group, princeton, problem, isbn, her, quantum hi, comput, vol, moscow, geometr, proof, prove, number, pp, s2cid
3	hi, speci, collect, that, her, had, museum, bird, natur, plant it, him, she, expedit, specimen, but, thi, be, botan, naturalist fossil, und, geolog, which, medic, british, would, insect, zoolog, or
4	that, astronom, translat, arab, it, hi, greek, astronomi, latin, philosoph hadith, him, god, be, treatis, thi, have, or, text, commentari but, had, write, are, caliph, all, would, observ, which, astrolog
8	genet, evolut, statist, that, bird, she, evolutionari, isbn, her, natur hi, human, bbc, it, be, anim, speci, popul, had, harvard zoolog, ornitholog, but, cambridg, would, thi, british, have, ha, about

Table 4.4: Top 30 words of focus communities based on TF-IDF scores. Scores from the remaining communities are seen from Appendix C.1.

Based on these 30 top words, we have given each of the focus communities a title, to sum up the main topics of that community. The titles are found by using the online large language model Chat GPT, to sum up the essence of the words. The titles are further boiled down to a name, which we will use to reference the given community later on. Titles and short names are as follows:

Community 1: Physics and Chemistry: Quantum, Nuclear, Particles, Energy, Mechanics (Physics)

Community 2: Mathematics: Algebra, Geometry, Topology (Math)

Community 3: Natural History: Specimens, Expeditions, Botany, Zoology, Geology, Fossils (NatHist)

Community 4: Astronomy and Philosophy: Greek, Arabic, Latin (Astronomy)

Community 8: Evolutionary Biology: Genetics, Species, Populations, Birds, Zoology, Statistics (Bio)

From the TF-IDF-analysis, we can conclude that the communities are indeed different in content. While they contained many Physics, Chemistry and Mathematics related words, Community 1 also contained words such as Soviet and Cambridge, and Community 2 contained space, quantum, and Moscow. Community 3 includes, besides the many natural history-related words, the words ”British” and ”und”, indicating both a British and German influence. Looking at community 4, we called this community ”Astronomy” and it contains words related to Arab and Greek astronomy and philosophy as well. The final focus community, community 8, also contains the word ”British” besides words related to biology and statistics.

4.5.6 Movement over time

To investigate how the communities develop over time, we have created the focus communities for 5 different years: 2008, 2012, 2016, 2020 and 2024 and plotted the movement between communities for the 5 selected focus communities from 2024 as a Sankey Diagram as seen from Figure 4.10. All communities smaller than 50 persons are allocated to the community ”Other” in order to simplify the figure where, whereas persons that were not yet in the network are in the group ”Not in data”.

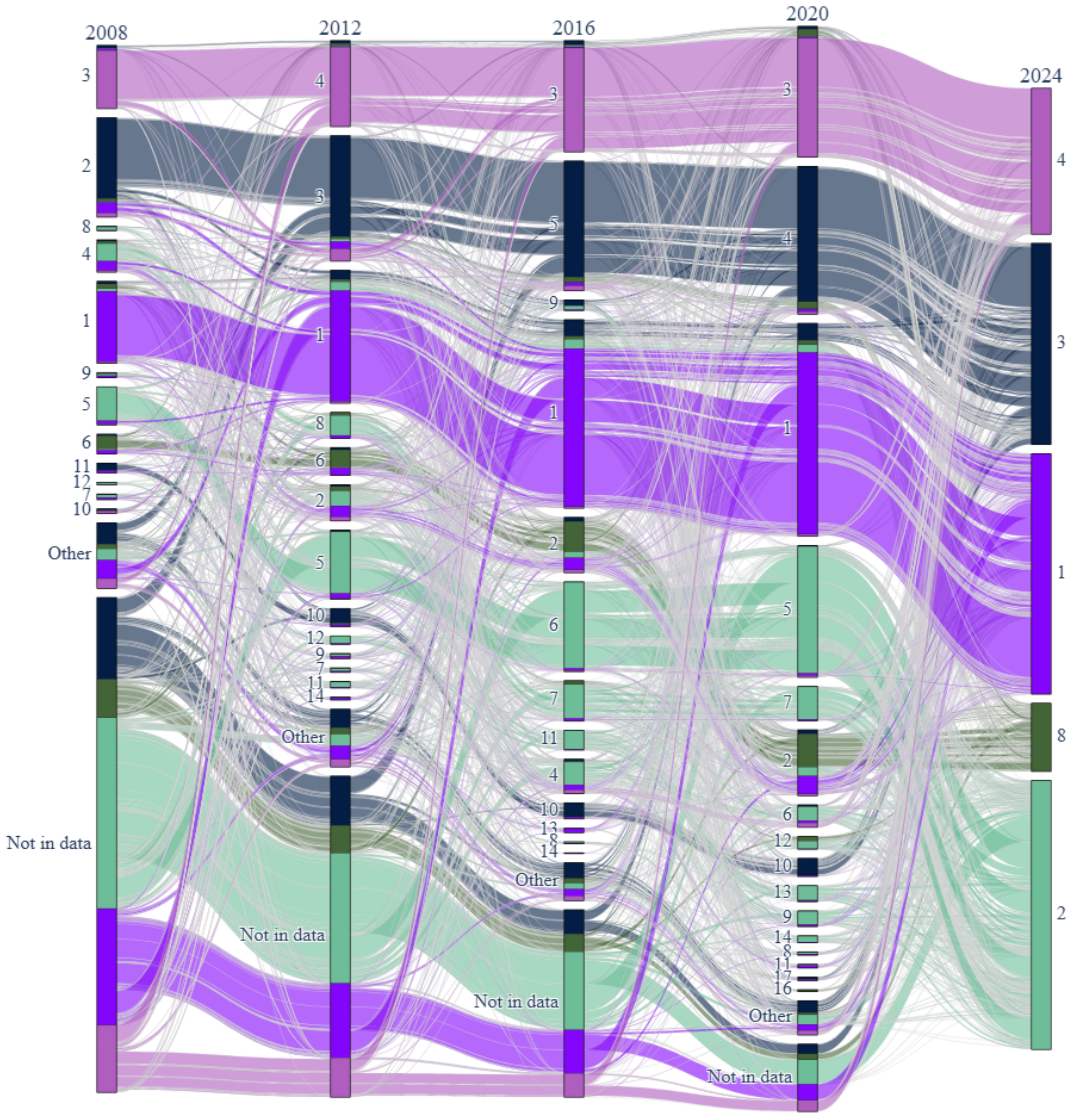


Figure 4.10: Movement of focus community members (2008-2024) coloured by 2024-communities.

We see that a large amount of the nodes in community 2 come from the artificially created community ”Not in data” in 2008 and 2012, which means that the articles have not yet been created at this time. This indicates that Community 2 is a young community compared to other communities such as 1, 3 and 4 that have members from what in 2008 was Community 1, 2 and 3.
Based on our knowledge about community 4, we expected this to have the largest amount of members

that already existed in 2008. This seems to be somewhat correct, but Community 3 also has a large amount of members who had an article in 2008.
In general, we see that the members of the previous communities do not seem to split up that often, but end up together in the same communities in 2024. This indicates that the links between the persons in the network have not changed a lot over time.

4.6 Network Growth

4.6.1 Bursty Growth over years

Having done an initial analysis of our network and investigated its structure, we will proceed to look into how the network grows over time. Specifically, we want to investigate whether the network exhibits bursty growth or steady growth over the years.

We define Bursty Growth as when new out-edges are added to nodes which are nearby nodes that recently gained out-edges. In order for it to be considered bursty, the growth has to happen within the same neighbourhood of the first growing node. We consider a neighbourhood as a node and the nodes that it shares in and out-edges with. In our small network in Figure 4.11, an example of a neighbourhood for the node $V2$ is its parent node $V1$ and its child nodes $V3, V4$ and $V5$.

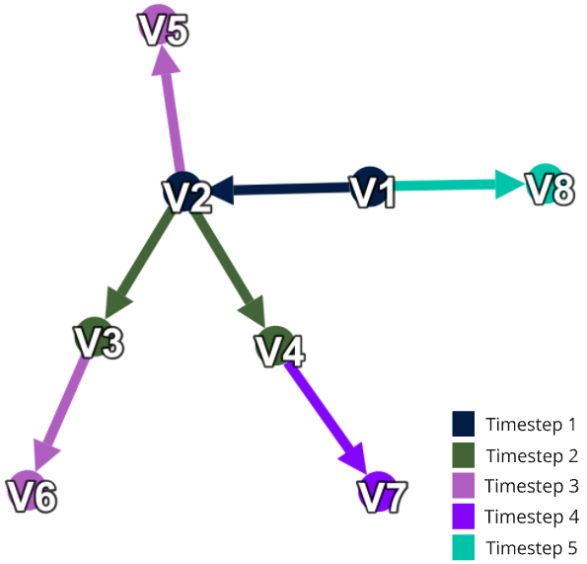


Figure 4.11: A small network example, consisting of 8 nodes, added over 5 time steps. The colour of nodes and edges indicates time steps. $(V1, V2)$ was added in t_1, $(V2, V3)$ and $(V2, V4)$ was added in t_2, $(V2, V5)$ and $(V3, V6)$ were added in t_3, $(V4, V7)$ was added in t_4, and $(V1, V8)$ was added in t_5

Using our definition of bursty growth, we can detect bursts in the small network in Figure 4.11 as an example. When $V1$ grows an out-edge in the first time step and its neighbour $V2$ grows out-edges the time step after, it would be considered bursty growth. Similarly, when $V2$ grows an out-edge in the third time step and its neighbour $V4$ grows an out-edge the time step after it is also considered bursty growth, even though the second edge isn't an extension of the growth from $V2$. Due to bursty growth happening over two time periods, we cannot detect any bursts in the last time period when $V1$ grows the last out-edge.

In order to test if our network grows more bursty than expected, we set up a permutation test, which is also a non-parametric statistical test. It works by reshuffling the data and creating a null-distribution which then can be compared to an empirical value [49]. In order to do so, we created a null hypothesis and a null model of what would be the expected growth. Firstly, the null hypothesis is formulated as follows

$$H_0 : \text{Nodes in the same neighbourhood are not gaining out-edges sequentially after each other more often than if the order of nodes gaining out-edges were randomly rearranged} \tag{4.9}$$

The expected growth would be growth where the probability that a node receives a new out-edge at time t is proportional to the gap between the out-degree of the node at t and at the end time T, over the sum of these gaps for all nodes in the graph. This means that if a node has already received an out-edge at $t = 1$, the probability of the same node gaining an edge at $t = 2$ is lower than at $t = 1$, making the probability of gaining an out-edge dependant on the node's own out-degree over time. This can be described with the following equation:

$$P_t(v \text{ gets next out-edge}) = \frac{d_T^{out}(v) - d_t^{out}(v)}{\sum_{w \in Graph(t)} [d_T^{out}(w) - d_t^{out}(w)]} \tag{4.10}$$

Applying this logic to the neighbourhoods in the network, if a node in a neighbourhood gained an out-edge in the last time step the neighbourhood is less likely to gain an out-edge in the current time step [50].

$$P_t(\text{Node in neighbourhood gets next out-edge}) = \sum_{v \in \text{Neighbourhood}} P_t(v \text{ gets next out-edge}) \tag{4.11}$$

Before implementing the bursty growth framework and the null model, we had to make a few assumptions regarding the growth in the network due to the way we scraped data and constructed our network. Firstly, our network consists of a data snapshot for each year, meaning that the network grows once a year. But at each of those 24 time steps, the network grows multiple out-edges and we can not identify which nodes grew them before others. This means that we are not able to identify the bursty growth that potentially happens within the specific year, and we expect that there could be more bursty growth in the network that we have not been able to identify. This is something that we will explore by investigating if nodes tend to have more out-edges added each year.
Since the network growth happens over several years and it is growing multiple out-edges each year, we also assume that it is possible for multiple cases of bursty growth to happen within the same time step in the network.

Taking these assumptions into account, we constructed the framework for detecting bursty growth. This is used to find the empirical count of bursts in our network. We started by investigating whether a node grows or not in the current year. Then, for the specific node, we identified the neighbourhood of the node the following year and investigated whether it also grew or not. If it did grow, it is counted as a burst. We applied the framework directly to the network and used it to count the empirical burst value.
Then, we implemented the null model. Iterating over the years, we sampled nodes where we calculated the probability of them gaining an edge using Equation . Here, we ensured that we added the same number of edges each year that was added to the empirical network so that the null-model networks and empirical network had the same size at each step. Then, for the sampled nodes, we define the neighbourhood for the next year and calculate the probability of adding an edge to a node in this neighbourhood. Finally, we draw a random number from a uniform distribution [0, 1], and if the random number is less than the calculated probability, we count it. We did this 1000 times to find a null distribution for the expected growth.

4.6.2 Growth rates

Having investigated the bursty growth in the graph, we also want to identify where in the graph growth happens throughout the years.
To get an idea of the growth across time, we use growth rate as a measure. This is calculated for each time step t, using the number of nodes for that year, $\mathbf{n}_t$, and for the former year, $\mathbf{n}_{t-1}$. Then, using a recursive approach, the growth rate can be calculated as the change in size for the sets of nodes compared to the former set of nodes.

$$NodeGrowth_t = \frac{\mathbf{n}_t - \mathbf{n}_{t-1}}{\mathbf{n}_{t-1}} \tag{4.12}$$

The same measure can also be calculated for the edges, using the collection of edges throughout the years, $\mathbf{e}_t$.

$$EdgeGrowth_t = \frac{\mathbf{e}_t - \mathbf{e}_{t-1}}{\mathbf{e}_{t-1}} \tag{4.13}$$

Using these measures, we calculate the node and edge growth rates for the entire graph, as well as for each of the communities that has more than 50 members. Then, we can compare the growth rate of the entire graph with the communities of the final graph to localise which sections of the graph grow faster than the entire graph as a whole.

4.7 Network Growth Results

4.7.1 Bursty Growth over years

Before testing the network for bursty growth, we are curious about how the nodes' out-degrees develop over time, since we interpret Equation 4.6.1 as suggesting that the out-degree growth of a node over time will be more linear in the expected growth. To investigate this for our network, we experimented with creating an out-degree plot, that could visualise the expected out-degree and actual out-degree in one figure. The out-degree plots for the biologist and geologist Louis Agassiz and mathematician Emmy Noether are shown in Figure 4.12.

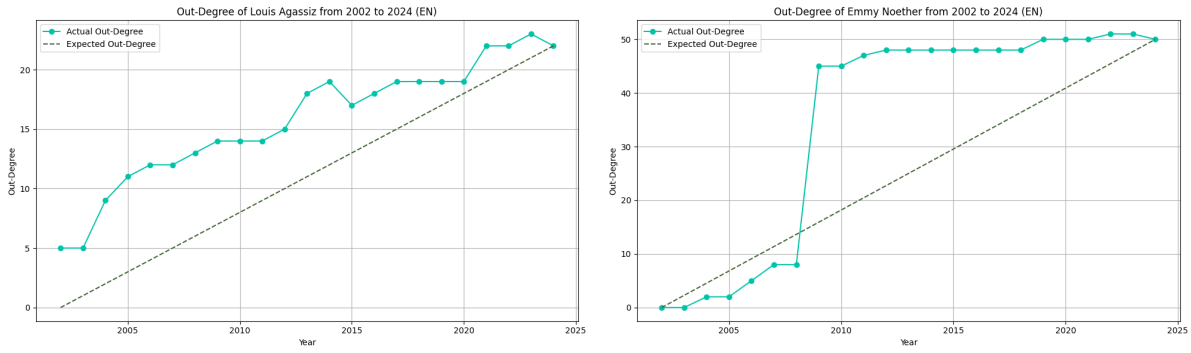


Figure 4.12: The difference between the node and edge growth rates for Louis Agassiz and Emmy Noether compared to growth rates for the entire network from 2006 to 2024.

A random sample of other nodes in the English network showed the same tendencies as Emmy Noether, where the out-degree growths more sporadically than evenly over time. On the other hand, Louis Agassiz is an example of a node which out-degree growths more evenly than most other nodes

including Emmy Noether.

The results of the permutation test for bursty growth are seen in Figure 4.13. Here, we see that the empirical value is much larger than the null distribution, which indicates bursty growth in our network. We have also investigated which of the years contained the most bursts and found that the years in the time period 2007-2008 to 2018-2019 contained the most, fluctuating only a bit between these years. This aligns with what we know about the network as we saw a large change in nodes and edges in 2007-2008 and the network continued having many edges added throughout the 2010s.

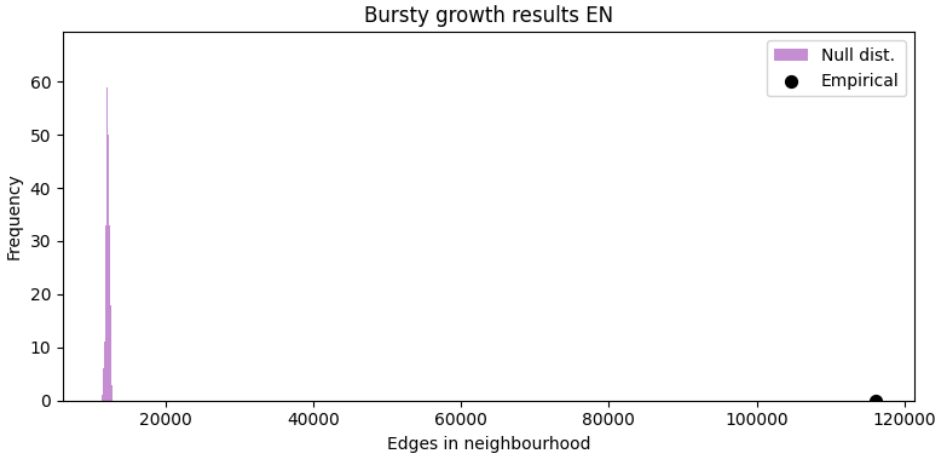


Figure 4.13: The empirical burst value and the expected burst values for the English network. The empirical values are located far from the null model, indicating bursty growth.

Because we knew in advance that our bursty growth detection would not take when growth happened to a node within the year into account, we also investigated the number of nodes growing multiple edges each year that we were potentially missing. We therefore counted the number of nodes each year that grew one out-edge and the number of nodes that grew more than one out-edge, which can be seen in Figure 4.14.

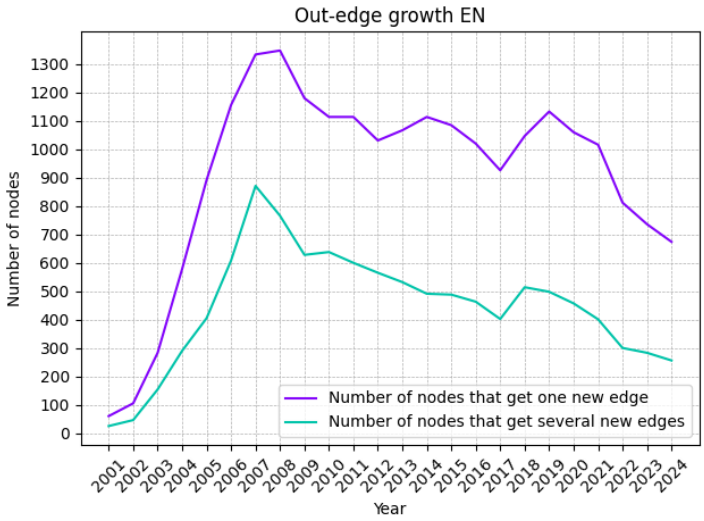


Figure 4.14: The number of nodes that each year grew one out-edge and the number of nodes that grew more than one out-edge.

While most of the nodes have one edge added each year, we see that the two trends have a pattern that follows each other. Up until 2021, the several-edge growing nodes make up 30-38% of all edge-growing nodes, while after 2021 they make up 27-29%.

4.7.2 Growth rates

The node and edge growth rates have been calculated for the entire English network, as well as for all communities that have more than 50 members. We have then found the difference between the growth rates for the communities and growth rates for the entire network so that when we plot them together, we can clearly see how much larger or smaller the growth rates of the communities are compared to the entire network. This plot can be found in Figure 4.15.

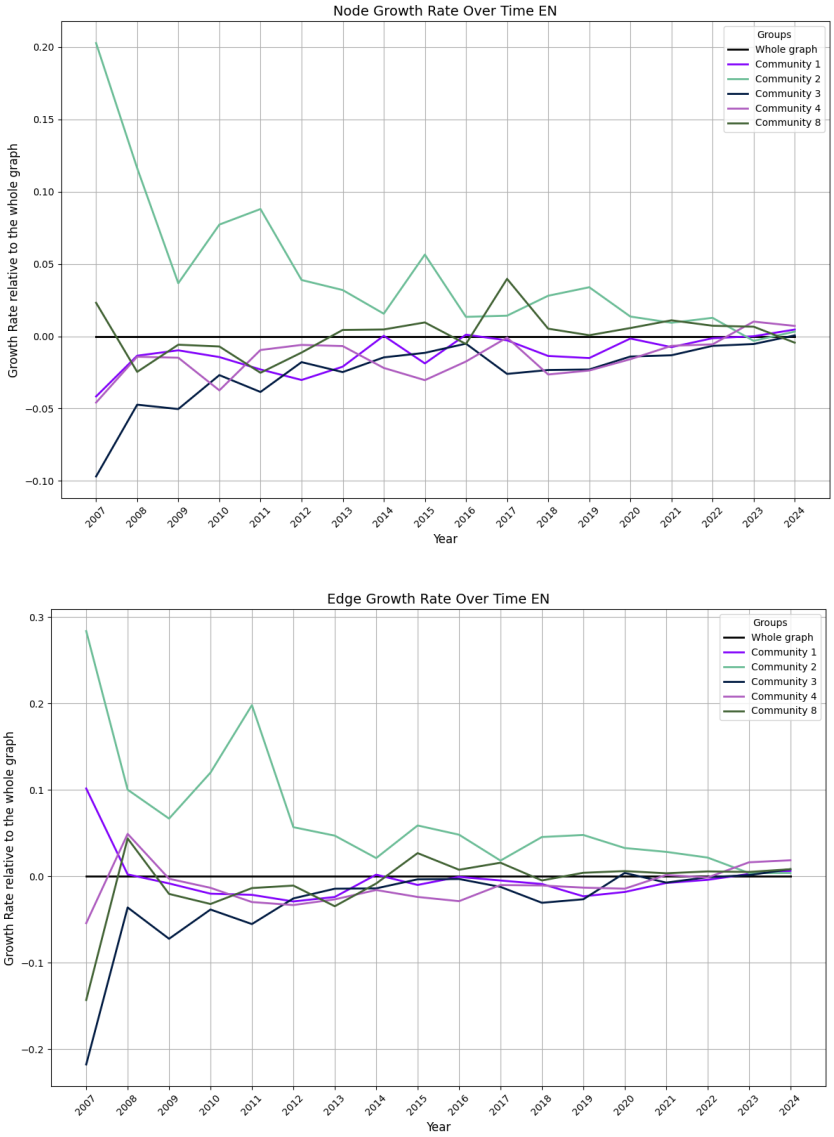


Figure 4.15: The difference between the node and edge growth rates for the English communities and the growth rates for the entire network from 2006 to 2024.

We see that both the edge and node growth rates in communities 1, 3 and 4 fall below the growth rate of the entire network from 2009 up to 2022, suggesting that the growth activity in the graph is higher in other parts of the network. Community 8 does have a larger peak in its node growth in 2017 but does otherwise stay around the growth of the entire graph.
The growth rates for community 2 is generally larger than the growth rates for the entire network and exhibits spikes in both node and edge growth rates in 2007, 2011 and 2015 compared to the other communities. This suggests that more growth activity is happening in community 2 compared to the other communities. We know from Figure 4.2 from Section 4.3 that the greatest amount of both nodes and links were added to the network in 2007, especially when considering the number of edges, which might be responsible for the first peak in the growth rates of community 2. The fluctuations of the edge and node growth rates are, in general, very large at the beginning of the time period. However, they all seem to decrease in width and stagnate, centring around the growth rate of the entire graph. This smoothness could indicate more stability in the network growth throughout the later years of the time period and that the growth is more evenly distributed across the different sections of the network as it gets larger and more connected.

CHAPTER 5

Comparison Across Languages

In the following chapter we change focus from looking entirely at the English network to including data from the German, French and Spanish versions of Wikipedia as well. We will compare the languages using the same methods for analysing the networks, their structure, the content of the articles and their growth. [1]

5.1 Data

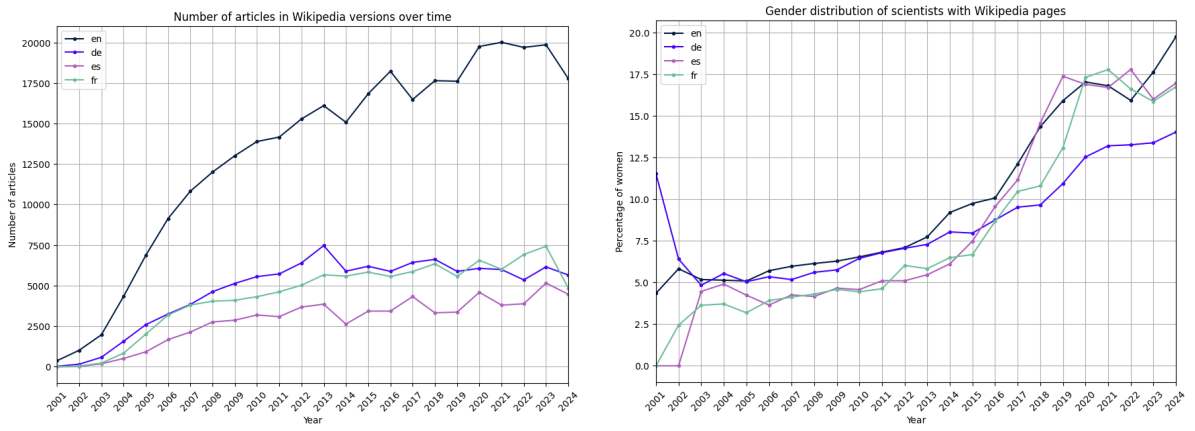
As for the English network, we have scraped revisions over time and the article contents from the list of scientists for the German, French and Spanish Wikipedia. In Table 5.1, we have collected information about the data from 2024 for each language, and we see that the English version is remarkably larger than the other three versions, with German being the second largest and Spanish being the smallest. We can further conclude that while the average birth year is within a close range for all four languages, there is a difference in the fraction of women. Where the English data consists of almost 20% women, the German is limited to only 14%.

Language	Average birth year	Percentage female	Number of articles found
English	1869.96	19.76%	17,772
German	1869.98	14.04%	5,655
French	1864.54	16.76%	4,851
Spanish	1855.87	16.98%	4,471

Table 5.1: Average birth year, percentage of female members and number of articles in data for 2024 data.

From Figure 5.1, the development in a number of articles as well as the percentage of female articles in the data for the 4 different language versions is visualized. The percentage of female articles is calculated as the number of articles that have been classified as women using our classifier divided by the total number of articles that year.

[1] All implementations and calculations for this chapter can be found in our GitHub repository: https://github.com/ElineBrunke/Gender-and-Language-Differences-in-the-Growth-of-Wikipedia.git. Here html-files for the Sankey plots as well as selected times series plots are found



(a) Number of articles **(b)** Gender distribution

Figure 5.1: Total number of articles and percentage of women in data over time. While the English graph is significantly larger than the other language versions, the four graphs seem to follow the same growth trends by growing at a high speed in the first 10 years and then stagnating in the last 10 years. The development in the gender distribution is similar for the English, French and Spanish graph while the German falls behind after 2015.

From Figure 5.1(a), the first thing to notice is the large difference in the number of articles between the English and the three other versions, as also mentioned above. It is also worth mentioning that it seems that the growth in the number of articles stagnated for the German, French and Spanish versions around 2013-2014. For the German version, there is even a small fall in the number of articles from 2013 to 2023, whereas the French and Spanish numbers continue to grow slightly in the same period. Looking at Figure 5.1(b) the overall proportion of women grows for all four languages after 2006 up until 2024. However, we see that after 2015 the English, French and Spanish proportion of women shows a much steeper growth compared to the earlier years. We also see that while the German version by 2024 has the lowest fraction of women, this has not always been the case. After 2015, we see a sudden increase in the fraction of female articles, but this tendency does not seem to have affected the German Wikipedia pages, as it is overtaken in growth by both the French and Spanish proportions of female articles.

5.2 Graph construction and visualisation

As we did for the English data, we have created 3 additional graphs for the German, French and Spanish data seen in Figure 5.2. These graphs are visualised and coloured by the focus communities we found in the previous chapter in order to get an indication of whether the communities from the English network also seem to group in our networks from other language versions.

(a) The German graph **(b)** The French graph **(c)** The Spanish graph

Figure 5.2: The German, French and Spanish networks, coloured by the English focus communities found in Section 4.5.3. There seem to be overlaps between the English communities and how the nodes connect in these graphs.

From Figure 5.2, we see that while the graphs are still large and difficult to comprehend, there seem to be some structures similar to those in the English graph. This indicates some similarities between the communities detected in Section 4.5.3 and the way the nodes are connected in the other languages. This is something we will investigate further when finding the communities for each of the other languages.

English	**German**	**French**	**Spanish**
Albert Einstein	Albert Einstein	Albert Einstein	Albert Einstein
Charles Darwin	David Hilbert	Charles Darwin	Charles Darwin
Aristotle	Charles Darwin	David Hilbert	Isaac Newton
Isaac Newton	Felix Klein	Felix Klein	David Hilbert
John von Neumann	Isaac Newton	Isaac Newton	Niels Bohr
Euclid	Paul Erdős	Henri Poincaré	Leonhard Euler
David Hilbert	Leonhard Euler	Paul Erdős	Felix Klein
Paul Erdős	Henri Poincaré	John von Neumann	John von Neumann
Niels Bohr	John von Neumann	Georges Cuvier	Georges Cuvier
Felix Klein	Alexander von Humboldt	Emmy Noether	Marie Curie

Table 5.2: Top 10 nodes in each network measured by in-degree

As we did for the English graph, the top 10 most connected nodes measured by in-degree in the German, French and Spanish graphs are found and seen in Table 5.2. The first thing we notice is that there are no women in the 10 most connected nodes in the English and German graphs. The Spanish and French each have one woman among the most connected nodes, respectively, Emmy Noether and Marie Curie. It is somewhat surprising that the German mathematician Emmy Noether is among the best connected in the French graph but not in the German and the Polish-French physicist and chemist Marie Curie is among the best connected in the Spanish graph but not in the French graph.

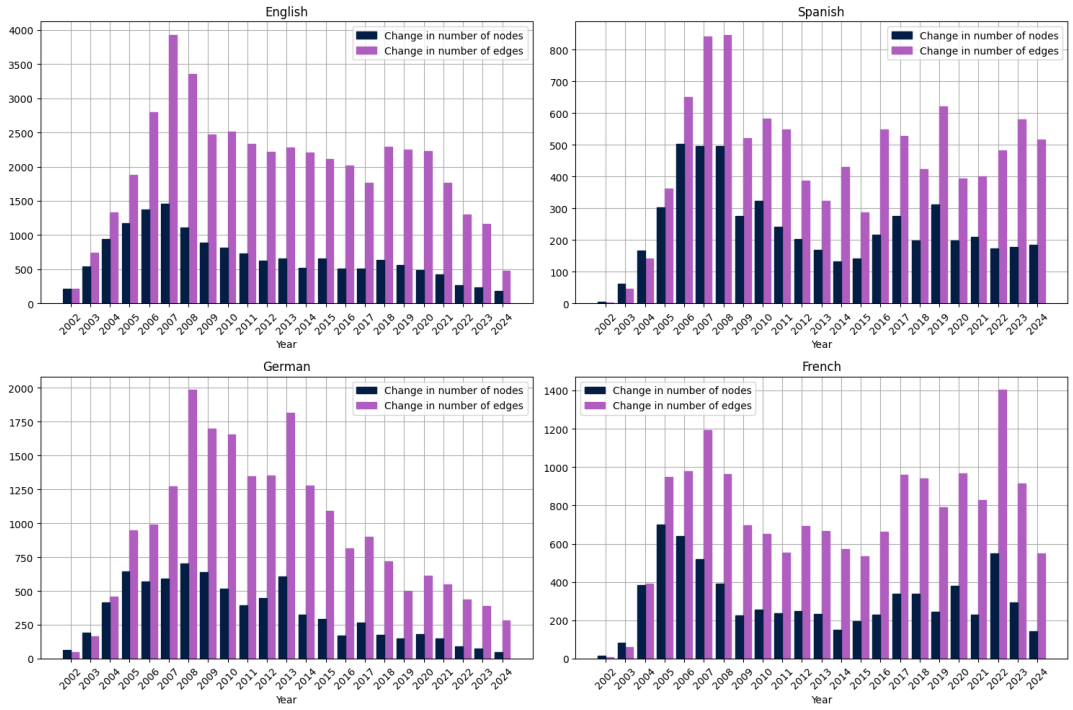


Figure 5.3: The development in number of nodes and edges over time. Each bar is the number of nodes or edges added that year.

From Figure 5.3, the development in a number of nodes and edges over time is seen. Where 2007 was the year with the largest growth for the English network in terms of both nodes and edges, a similar peak is seen in 2008 in the German graph. In the Spanish, the peak happened over several years from 2006 to 2008, and for the French, we see peaks in two different years, 2005 for the number of nodes and 2022 for the number of edges added. This is different from the other languages, where the growth seems to fall from around 2008 and forward.
All graphs grow more edges than nodes each year. However, the English and German networks seem to grow the most edges proportionally to nodes compared to the French and Spanish networks. This is investigated further in Figure 5.4, where the connectivity is measured as the number of edges per node.

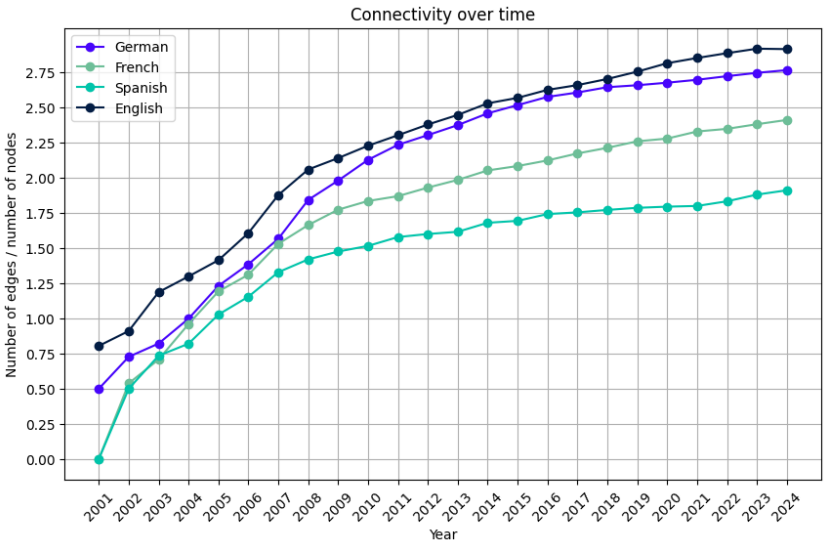


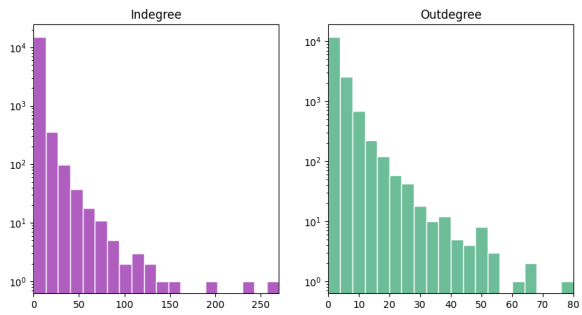
Figure 5.4: The connectivity of the four graphs measured over time, calculated as the number of edges per node. The Spanish is the least connected and the English is the best connected.

We see that while the English network is much larger, the German network is almost as well-connected as the English in 2024. From the beginning in 2001 to today the number of edges per node is growing, most rapid in the first 10-15 years, but still growing in a slower pace after 2013. This indicates that the network gets better an better connected over time which we expect will result in a more robust network.

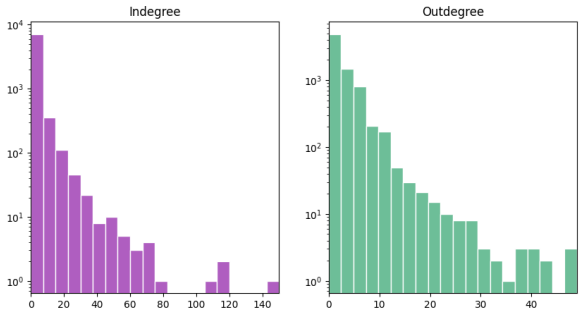
5.3 Network Structure

5.3.1 Degree

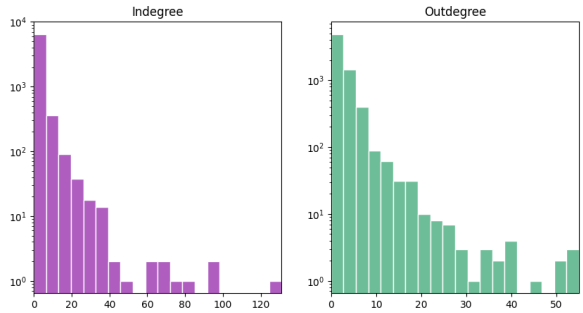
Similarly to the degree distribution for the male and female nodes created for the English network, we also created the in- and out-degree distribution for all four language versions, shown in Figure 5.5.



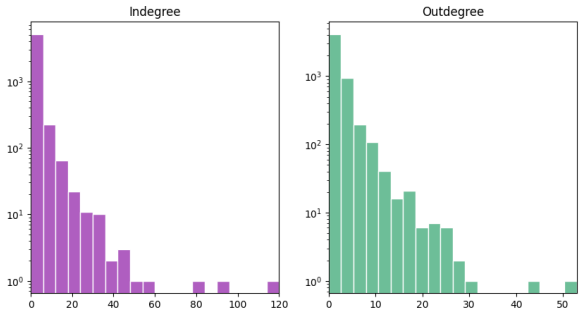
(a) The English degree distribution



(b) The German degree distribution



(c) The French degree distribution



(d) The Spanish degree distribution

Figure 5.5: Degree distribution split in in- and out-degree for all four languages, meaning the distribution of number of links from and to articles in the network. Note that the axis differs between the plots.

Similar for all of the languages is the out-degree distribution being more dense than the in-degree distribution. This indicates that nodes generally have fewer out-edges and tend to point to the same nodes in the graph. This is a pattern that indicates that the distributions follow a power-law pattern, which we previously covered for the English network, where a small number of nodes are highly connected and most other nodes have very few connections.
The English network has the widest range for both in-degree and out-degree, indicating that it is generally more densely connected than the other networks. The German, French and Spanish networks have shorter tails in their degree distributions, indicating fewer dense hubs with highly connected centres and less connectivity overall compared to the English network.
As we saw earlier in Figure 5.4, the German network had high connectivity, and out of the three other languages, it seems to have more hubs with highly connected centres. This indicates that it behaves closest to the English network. While the French network falls somewhere in the middle of the three, the Spanish network has fewer nodes with a high amount of outgoing links and seems to have more evenly distributed in and out degrees.

5.3.2 Robustness and k-core

As for the English network, we also found the k-cores for the graphs of the German, French and Spanish networks. We calculated the same statistics as we did for the English graph and it is presented in Figure 5.6.

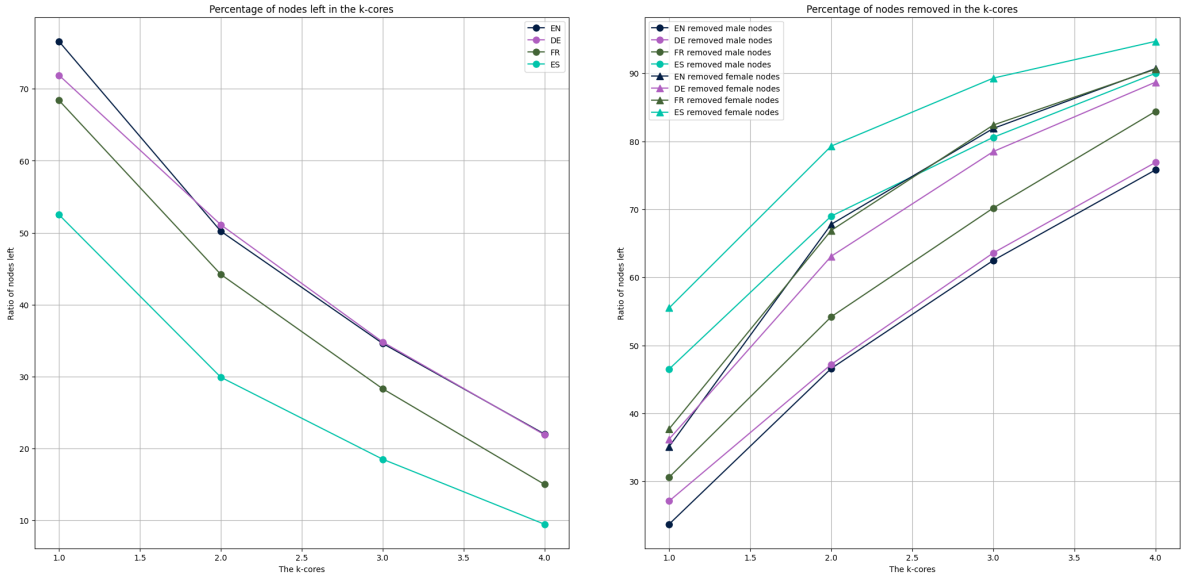


Figure 5.6: An overview of the calculated k-core statistics for the English, German, French and Spanish networks. The left figure shows the percentage of remaining nodes in each k-core for all the languages. The right figure shows the percentage of removed male and female nodes in each k-core for all the languages.

When comparing the percentage of nodes left in the k-cores, we see that the Spanish network loses a larger amount of nodes in each of the k-cores, showing that it has the most nodes located in the periphery. We also see while the German network has fewer nodes and links than the English, they both have around the same percentage of nodes left in each of their k-cores. This aligns with their similar connectivity in 2024 shown in Figure 5.4, suggesting that they have a similar amount of nodes located in the periphery.

We also see that the German network has a smaller percentage of their female nodes removed in its 2, 3 and 4-core. We know from Table 5.1 that the German network has the smallest share of female nodes overall, but from the k-cores, it is suggested that the German female nodes are better connected than the female nodes in the other three networks. However, the German network also has the overall lowest percentage of female nodes in the 4-core, having only 5.57%. It is slightly less than the Spanish network, whose 4-core has 5.79% female nodes left after removing the largest share of the four networks. The French network has the largest share of female nodes left at 8.37% but is also shown to have a larger number of male nodes removed in all k-cores than both the English and German networks. Similar for all four languages is that the robustness of the networks is mostly dependent on the male nodes.

5.3.3 Communities

Similar to what was done on the English network, we have created InfoMap communities for the German, Spanish and French networks. For the 2024 networks, this resulted in 304 German, 310 French and 369 Spanish communities. Filtering to only include communities larger than 50 members, we ended up with 14 German, 12 French and 14 Spanish communities. The number of members and percentage of female members of these communities is seen in Figure 5.7.

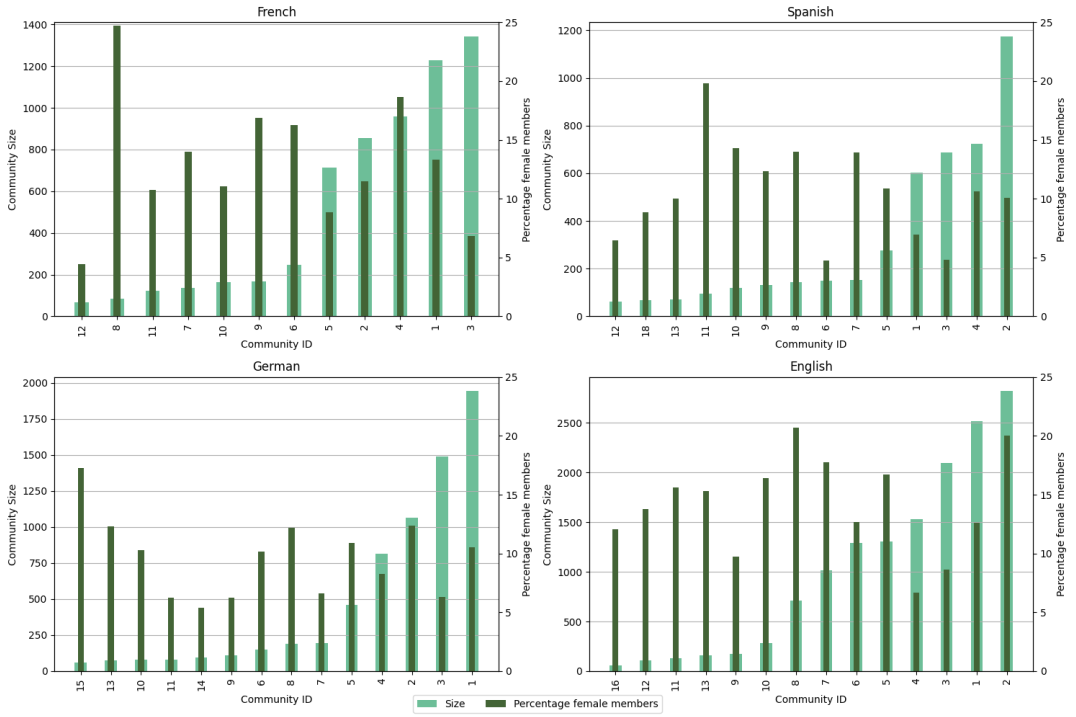


Figure 5.7: Percentage of female members and size of communities larger than 50 members for all four languages. The left axis shows the number of members in the community and the right axis is the percentage of female members

As previously concluded, we see that the German network, in general, has fewer female nodes, but we can also conclude that the percentage of female nodes in the largest communities, in general, is low for the German communities. In comparison, the French network has community 8 with almost 25 % female members, which is the highest share of female members out of all communities across the languages. This is, however, a rather small community with less than 100 members.

5.3.4 Community Strength

As for the English network, we have calculated the strength measure introduced in Section 4.4.4. The results are seen in Table 5.3.

German			French			Spanish		
Community	**Size**	**Strength**	**Community**	**Size**	**Strength**	**Community**	**Size**	**Strength**
1	1943	90.83%	1	1229	88.69%	1	603	93.20%
2	1066	86.20%	2	854	86.07%	2	1175	95.74%
3	1491	93.96%	3	1343	95.01%	3	687	93.60%
4	813	87.21%	4	959	87.80%	4	725	95.45%
5	460	79.35%	5	712	91.85%	5	276	94.20%
6	148	74.32%	6	246	71.54%	6	148	77.03%
7	197	81.73%	7	136	80.88%	7	151	87.33%
8	189	88.36%	8	85	77.65%	8	143	93.01%
9	112	96.43%	9	166	89.76%	9	130	88.46%
10	78	85.71%	10	163	71.78%	10	119	89.92%
11	80	78.75%	11	121	85.12%	11	96	93.75%
13	73	80.82%	12	68	89.71%	12	62	88.71%
14	93	88.17%				13	70	95.71%
15	58	77.59%				18	68	89.71%

Table 5.3: The community strength and size of all communities larger than 50 members for the German, French and Spanish graphs. The strength measure indicates the percentage of members in the community that are connected to more members in their own community than outside the community. A strength of 100 % indicates a strong community.

From the results, we see that communities 3, 9 and 1 are the 'strongest' in the German graph, meaning that the nodes of these communities tend to link more to nodes within the same community than nodes from other communities. For the French graph, communities 3 and 5 are the strongest connected and for the Spanish communities 1, 2, 3, 4, 5, 8, 11 and 13 all have a strength measure above 90%. This means that more than 90 % of the nodes in these communities have more edges to nodes in the same community than outside the community. None of the communities in the 3 graphs can be characterize as **a strong community** since this would need 100% of the nodes in a community to be more connected to their own community than others. However, we see that the Spanish communities tend to be *stronger* than the French and German.

5.3.5 Similar communities across languages

We want to know more about how the communities differ across languages and test if it is possible to map the communities to each other across languages. To do so, we have counted the number of identical community members in each combination of communities for each foreign language and English. This is done by creating 3 confusion matrices seen in Figures 5.8 and 5.9.

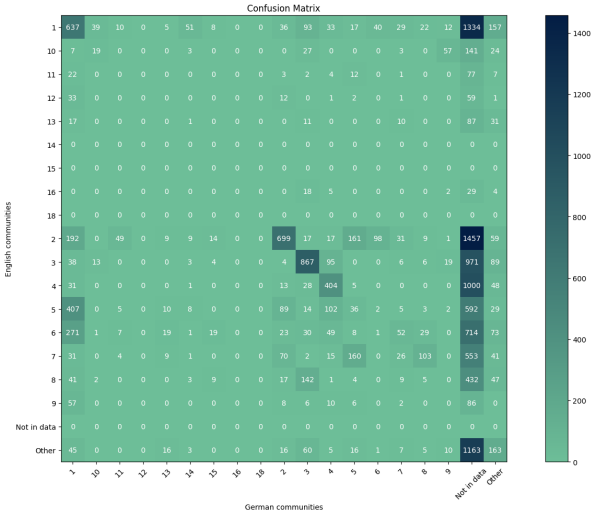


Figure 5.8: Confusion matrix of German and English communities. Each number is the number of people that is in that given combination of communities. For instance, 192 nodes in the English community 2 are members of the German community 1

From Figure 5.8 we can conclude that there are indeed overlaps between the German and English communities. Looking at the focus communities from the English network, we can see that these, to a large extent, correspond to communities in the German network. The same is the case for the French and Spanish communities seen in Figure 5.9. A mapping of the most similar communities across languages is seen from Table 5.4.

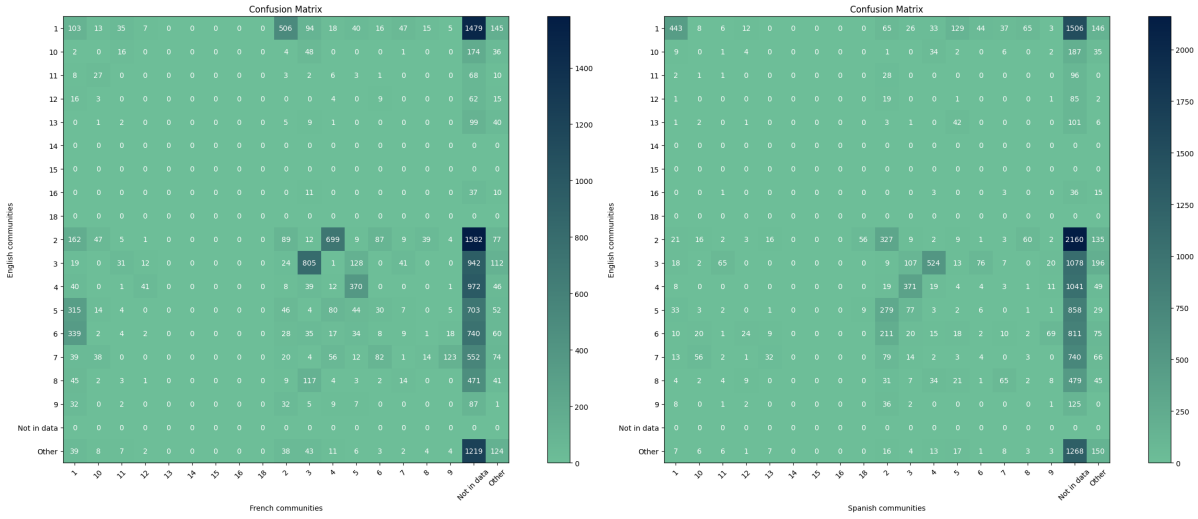


Figure 5.9: Confusion matrix of French and Spanish communities compared to the English communities.

As seen in Table 5.4, we can not only map the English communities with the German, French and Spanish, but we also detect an overlap between the English communities 3 and 8. These are both very similar to the German and French community 3 according to our mapping. This indicates that the two communities might have many connections to each other, which is also seen from the initial illustration of the graph in Figure 4.6. In this graph, Community 3 is coloured dark blue, and Community 8 has a

light teal colour. When looking at the plot, we see that these two communities are located right next to each other, also indicating many common edges.

English community ID	Most similar German community	Most similar French community	Most similar Spanish community
1	1	2	1
2	2	4	2
3	3	3	4
4	4	5	3
8	3	3	7

Table 5.4: The most similar communities for each English focus community based on confusion matrices. Please note that this is entirely based on the community with the most similar members. For some communities, there are fewer overlaps than others, for instance, the English community 8 and the Spanish community 7 has only 65 nodes in common and the English community 8 has similarities with other Spanish communities as well

To better understand the movement between communities across languages, three Sankey plots have been created and are seen in Figure 5.10.

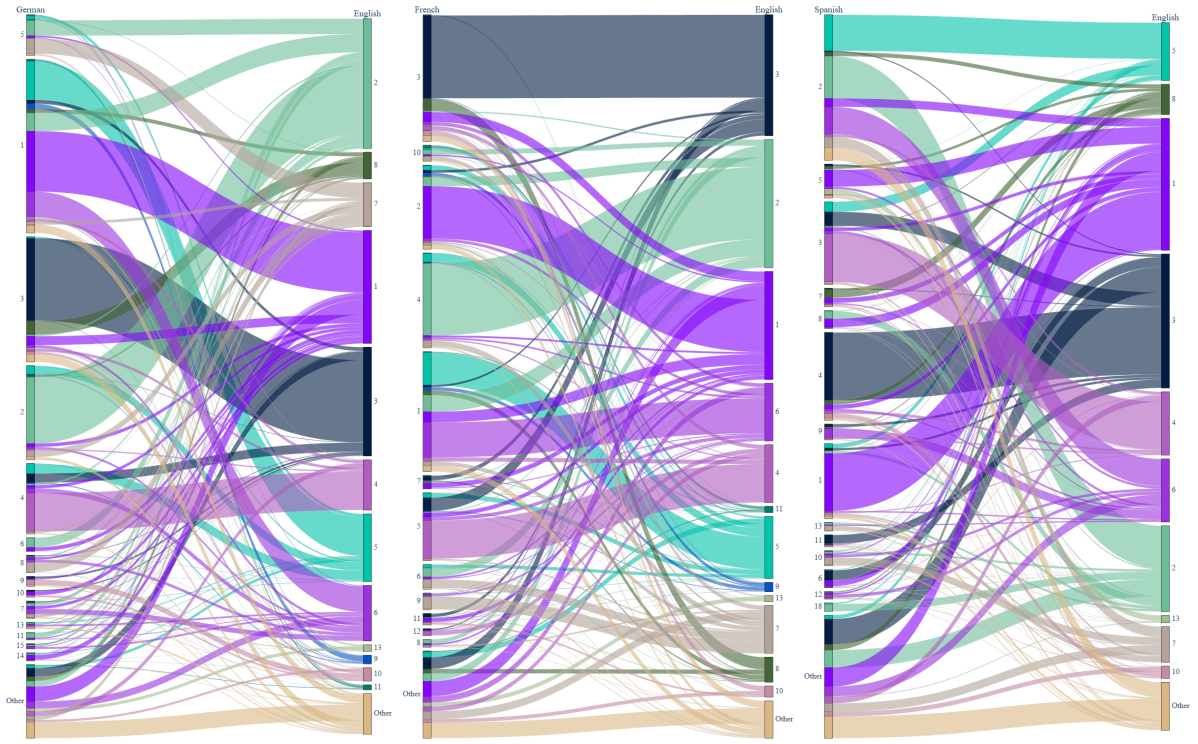


Figure 5.10: Comparison of communities across language. Communities smaller than 50 members are assigned to 'Other'. Persons who are in the English graph but not in the graph of the other language are not included in these plots.

From the Sankey plots, we can conclude that while there are several similarities, as we saw in the confusion matrices, there are also differences. Looking at the German community 1 this is split into several different communities in the English graph, and while it does have many overlapping members with the English community 1, it also shares many members with the English communities 2, 5, 6 and 9.

In our further analysis of the communities, we will look at the communities seen in Table 5.4. This is chosen as these are the communities that, in numbers, overlap the most with the English communities.

5.3.6 Most connected nodes

The top 10 most connected nodes of each of the previously selected communities are seen in Table 5.5. The degree is measured by in-degree.

German	French	Spanish
Community 1 Albert Einstein, David Hilbert, Felix Klein, John von Neumann, Arnold Sommerfeld, Hermann Weyl, Niels Bohr, Richard Courant, Emmy Noether, Max Planck	**Community 2** Albert Einstein, Henri Poincaré, Niels Bohr, Marie Curie, Werner Heisenberg, Max Planck, Hermann Weyl, Paul Langevin, Enrico Fermi, Jacques Hadamard	**Community 1** Albert Einstein, Niels Bohr, Marie Curie, Enrico Fermi, Max Born, Werner Heisenberg, Ernest Rutherford, Arnold Sommerfeld, Max Planck, Henri Poincaré
Community 2 Henri Poincaré, Alexander Grothendieck, André Weil, Barry Mazur, Laurent Schwartz, Michael Atiyah, Jean-Pierre Serre, Jean Dieudonné, Charles Hermite, Henri Cartan	**Community 3** Charles Darwin, Georges Cuvier, Alexander von Humboldt, Louis Agassiz, Louis Pasteur, Richard Dawkins, Karl Pearson, David Starr Jordan Joseph Banks, Hermann von Helmholtz	**Community 2** David Hilbert, Felix Klein, John von Neumann, Bertrand Russell, Emmy Noether, Alan Turing, Kurt Gödel, Paul Erdős, Karl Popper, Ludwig Wittgenstein
Community 3 Charles Darwin, Alexander von Humboldt, Georges Cuvier, Louis Agassiz, Johann Wolfgang von Goethe, Rudolf Virchow, Erwin Stresemann, Charles Lyell, Michael Faraday, Ernst Haeckel	**Community 4** Henri Cartan, Jean-Pierre Serre, André Weil, Emil Artin, Barry Mazur, Alain Connes, Michael Atiyah, Andrew Wiles, Pierre Deligne, Claude Chevalley	**Community 3** Isaac Newton, Leonhard Euler, Johannes Kepler, Galileo Galilei, Carl Friedrich Gauss, René Descartes, Tycho Brahe, Blaise Pascal, Benjamin Franklin, Christiaan Huygens
Community 4 Isaac Newton, Leonhard Euler, Moritz Cantor, Galileo Galilei, Gottfried Wilhelm Leibniz, René Descartes, Leopold Kronecker, Johannes Kepler, Archimedes, Joseph-Louis Lagrange	**Community 5** Isaac Newton, Leonhard Euler, François Viète, René Descartes, François Arago, Tycho Brahe, Gottfried Wilhelm Leibniz, Joseph-Louis Lagrange, Johannes Kepler, Blaise Pascal	**Community 4** Charles Darwin, Georges Cuvier, Charles Lyell, Louis Agassiz, Joseph Banks, Charles Babbage, Karl Pearson, Michael Faraday, Augustus De Morgan, Ernst Haeckel
		Community 7 Richard Dawkins, Konrad Lorenz, Stephen Jay Gould, Noam Chomsky, Ernst Mayr, John Searle, Steven Pinker, Daniel Dennett, Albert Claude, Christian de Duve

Table 5.5: The most connected nodes in each of the focus communities measured by in-degree. We see many overlaps in the most connected nodes across languages.

The first thing we notice when looking at the most connected nodes in the German focus communities is that Emmy Noether is now among the top 10 in community 1. However, she is not in the corresponding French community 2, though she was present in the overall top 10 most connected in the French network. Looking her up, she appears in the French community 1, which is not one of our focus communities. Emmy Noether is a good example of the difference in communities across languages. In the German network, she is in the same community as Albert Einstein and Niels Bohr, but in the Spanish network, she is in another community. In other words, we can conclude that while the mapped communities do have large overlaps with the same English communities, it is not only the smaller nodes that move between communities, but also some of the most connected nodes.

5.3.7 Content

To gain more insights about the communities we have calculated the TF-IDF scores for the English versions of the articles present in each language network. The top 30 words of each community are seen in Appendix C.2, C.3 and C.4. As for the English TF-IDF results, we have summarized the content of our focus communities using an LLM. The following are the titles for the focus communities:

German 1: Science and Philosophy: Quantum, Logic, Nuclear, Energy, Algebra (Physics)

German 2: Mathematics: Algebra, Topology, Geometry, Functions, Cohomology, Proofs (Math)

German 3: Natural History: Species, Birds, Plants, Zoology, Fossils, Expeditions, Museums (NatHist)

German 4: Astronomy and Philosophy: Observations, Motion, Translations, Treatises (Astronomy)

French 2: Quantum Physics: Atoms, Particles, Energy, Mechanics, Radiation, Computation (Physics)

French 3: Natural History: Specimens, Evolution, Fossils, Zoology, Botany, Ornithology (NatHist)

French 4: Mathematics: Geometry, Algebra, Topology, Manifolds, Equations, Cohomology (Math)

French 5: Astronomy and Philosophy: Stars, Motion, Observations,Translations (Astronomy)

Spanish 1: Nuclear Physics: Atoms, Energy, Radiation, Reactors, Particles, Experiments (Physics)

Spanish 2: Mathematics and Philosophy: Logic, Algebra, Geometry, Functions, Economics (Math)

Spanish 3: Astronomy and Philosophy: Latin, Motion, Observations, Treatises, Light (Astronomy)

Spanish 4: Natural History: Plants, Animals, Fossils, Specimens, Expeditions, Botany, Zoology (NatHist)

Spanish 7: Evolution and Cognition: Genetics, Species, Humans, Behavior, Language (Bio)

We see that the communities we detected as similar, also have similarities when looking at the most significant words of the communities. The short titles for each focus community are seen in Table 5.6

English	**German**	**French**	**Spanish**
Physics	Physics	Physics	Physics
Math	Math	Math	Math
NatHist	NatHist	NatHist	NatHist
Astronomy	Astronomy	Astronomy	Astronomy
Bio	NatHist	NatHist	Bio

Table 5.6: Short titles of each community for each language version mapped after confusion-matrix-similarities.

We see that the content of the communities we have mapped is similar, especially for the Astronomy and Natural History communities. The English and Spanish communities do however distinguish the Natural History community in two; Natural History and Bio. Compressing the content of a least 50 Wikipedia articles down to one word naturally leaves out many nuances and differences between the communities across language, but we see that the overall topics of our mapped communities are the same with the one difference that some languages split Bio and NatHist in two.

5.4 Network Growth

5.4.1 Bursty Growth

Having investigated the differences in structure across all 4 networks, we will move on to exploring the growth patterns, starting with the bursty growth detection. The same permutation test that was performed on the English network was also performed on the German, French and Spanish networks, and the results can be found in Figure 5.13.

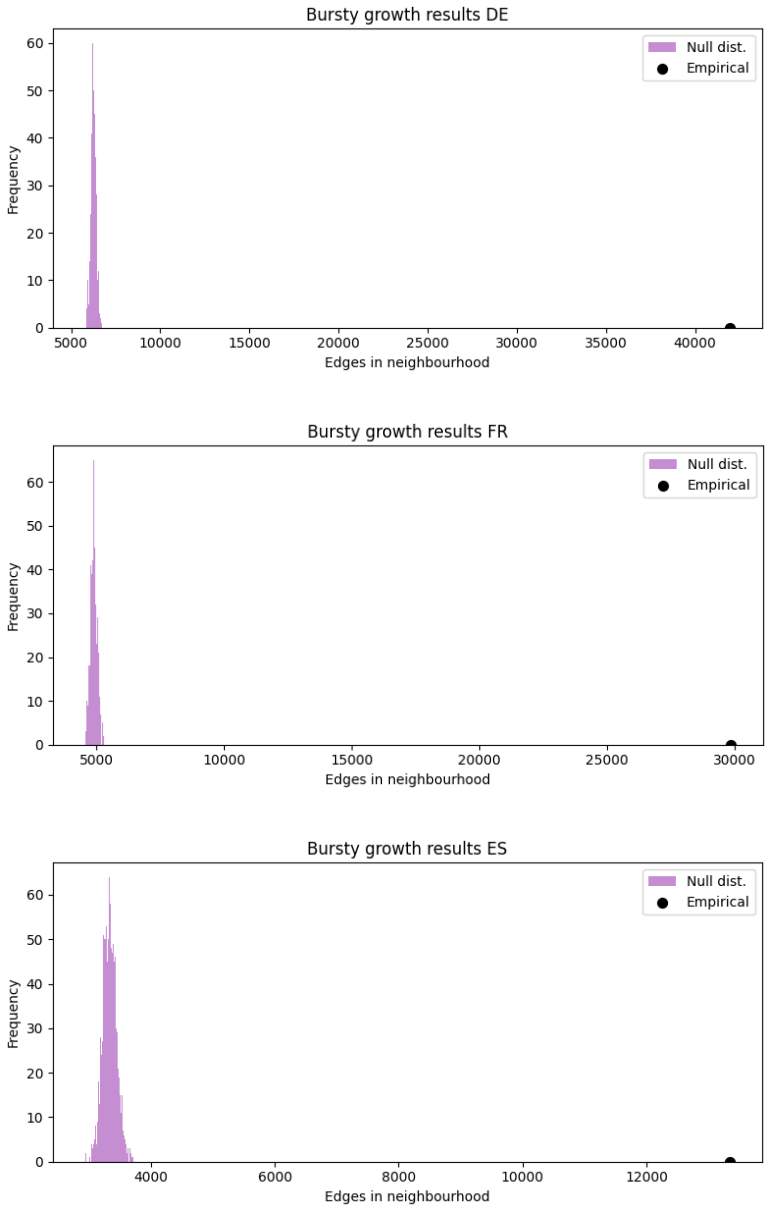
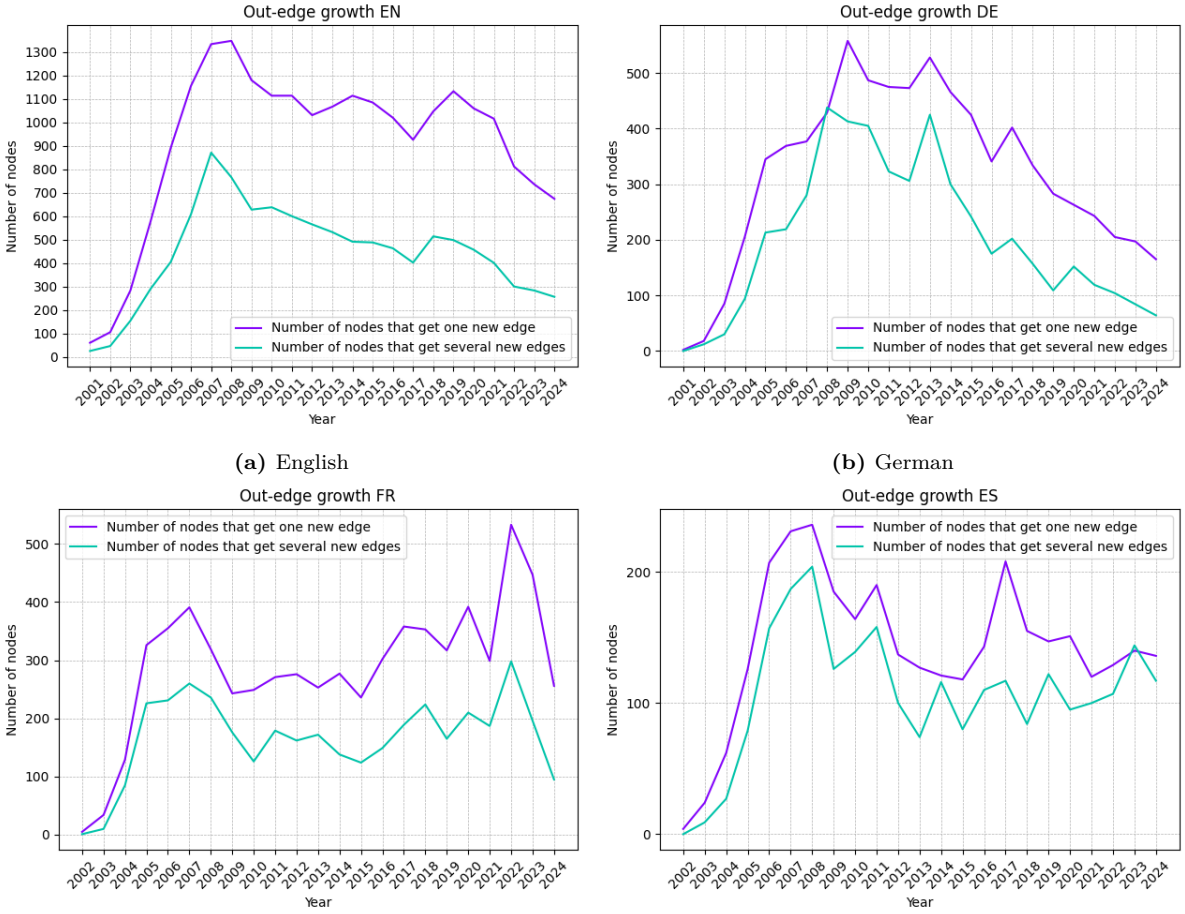


Figure 5.11: The permutation test results, with the empirical burst value and the expected burst values for the German, French and Spanish networks. All three networks exhibit bursty growth.

For all three networks, we see that the empirical values are larger than the expected growth, indicating that the network growth is bursty. However, compared to the English permutation test in Figure 4.13, the distance between the expected growth and empirical values for these networks is not as stark of a contrast, which implies that the English Network is more bursty. Judging on the width of the null distributions and distance to the empirical value, the German network would be the second most bursty, followed by the French and then the Spanish network.
We also investigated which years were the most bursty for the three networks by counting the number of bursts they experienced each year. Here, we found that the German network contained the most bursts in 2008-2009 to 2012-2013, after which its number of bursts continued to decline. For the French network, the number of bursts increases until 2020-2021, while for the Spanish network, it fluctuates

more, having peaks in 2007-2008, 2015-2016 and 2021-2022.

We also investigated the number of nodes growing multiple edges each year for the German, French and Spanish networks. The results is shown in Figure 5.12, along with the results for the English network.



(a) English **(b)** German

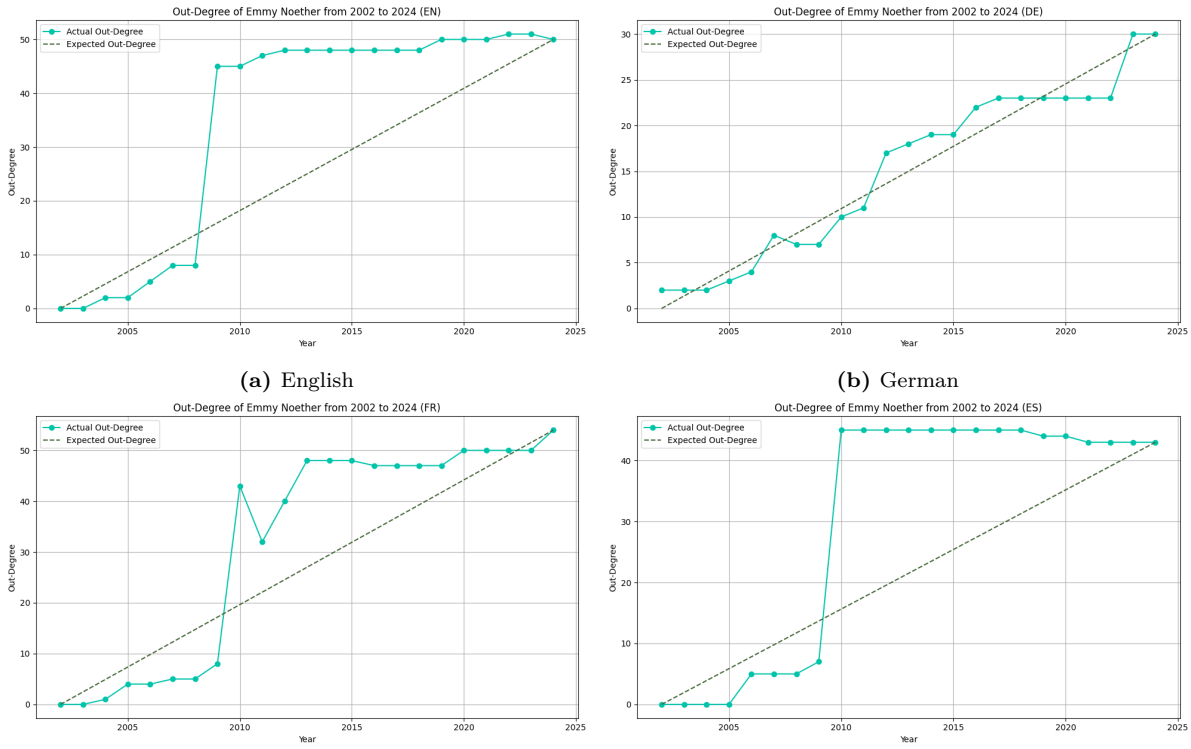
(c) French **(d)** Spanish

Figure 5.12: The number of nodes that each year grew one out-edge and the number of nodes that grew more than one out-edge. For all four languages, nodes tend to only grow one edge a year, but for the German and Spanish, the two lines are close, indicating that almost half of the nodes that grew nodes grew more than one.

We see that as for the English network, the networks of the other languages have similar patterns when it comes to the nodes growing multiple edges and nodes only growing one edge per year. However, the stark difference when comparing them is that the trend lines for the German, French and Spanish networks are much closer together than for the English network. We see that where this difference for the English network was around 27-38%, it is overall closer to 40-50% for the three other networks. It seems that the number of nodes growing multiple out-edges per year is proportionally much higher than for the English graph.

Since our four networks contain many of the same scientists, it is interesting to investigate if individual nodes grow differently across languages. In order to explore if the nodes grows the same way, we

have investigated the out-degrees of the same nodes in the different networks, using the same out-degree plot from Figure 4.12.



(a) English **(b)** German

(c) French **(d)** Spanish

Figure 5.13: The Out-degree plots for Emmy Noether for all four languages. We see that while the graph grows similar across language, the individual growth for the same person can differ a lot across language.

The stabilisation of the out-degree growth suggests that after a period of more rapid growth, there was editing activity on the article and fewer revisions for the article in the English, French and Spanish networks. This can, however, not be said about the node in the German network, which experiences growth in out-degree that is very close to the expected.
This means that the even though the node is present in all networks, it does not grow out-edges in the same pattern across the networks.

5.4.2 Growth Rates

The node and edge growth rates have been calculated for the entire German, French and Spanish networks, as well as for all their communities that has more than 50 members. As done previously we have calculated the difference between the growth rates for the communities and growth rates for the entire network, so that when we plot them together, we can clearly see how growth rates of the communities are compared to the entire network.

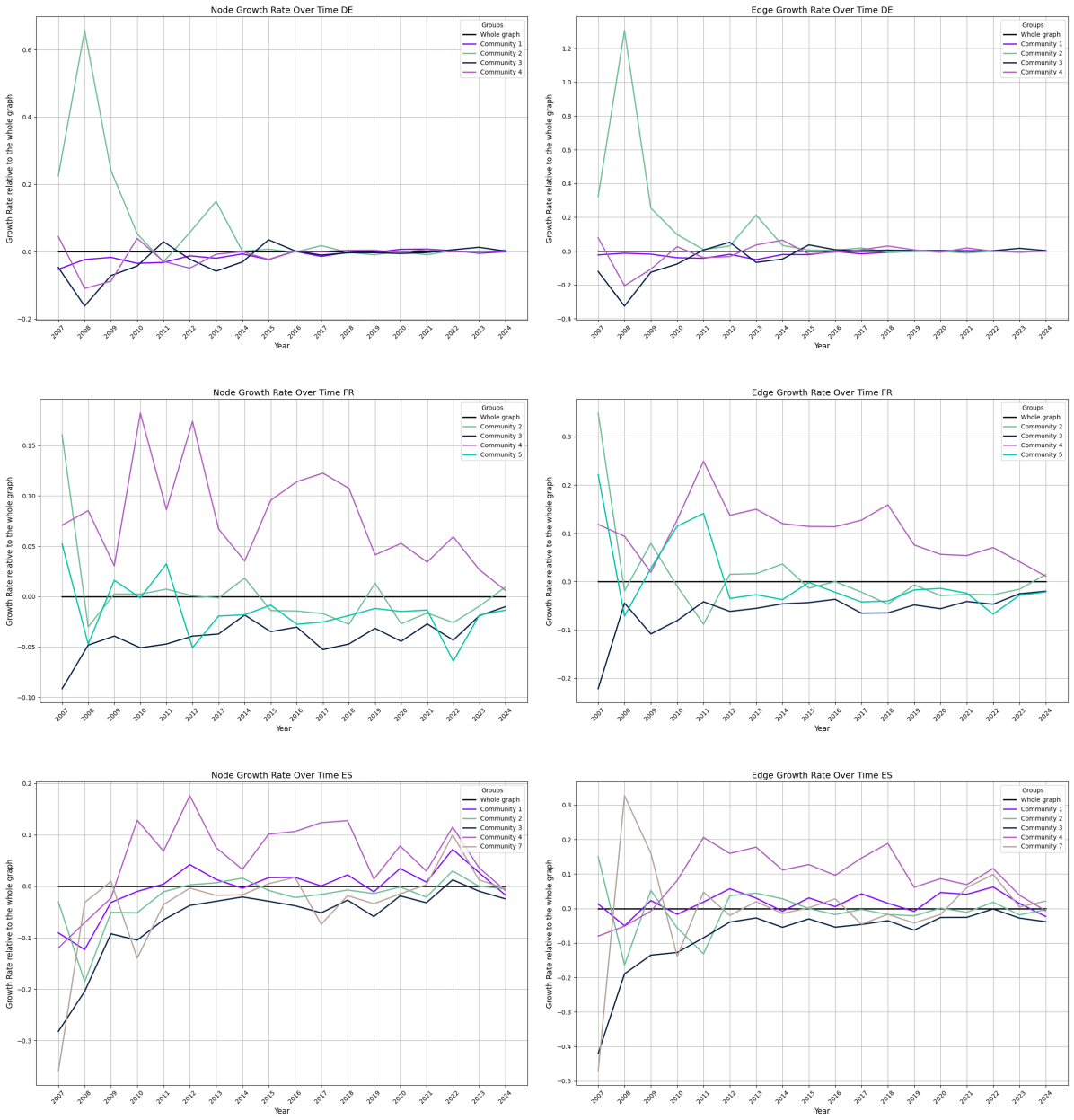


Figure 5.14: The difference between the node and edge growth rates for the German, French and Spanish communities compared to the growth rates for the entire network from 2006 to 2024.

We do see some of the same similarities in the community growth rates between the languages. Similar to what we saw for the English network, the German and French networks experience one community having overall higher growth rates compared to the rest of the networks. For the German network it is community 2, which we earlier identified as a Math network and for the French network it is community 4, which we also identified as a Math network, just like the English community 2.
This behaviour is not seen in the Spanish network, where we instead see the NatHist and, later also, the physics community having a higher growth rate than the rest of the network.
The German network displays a peak for Community 2 in 2008, which the English Community 2 had around 2007. However, the peak in for the German community 2 is much more drastic and could be

related to the big change in edges, which is also seen in Figure 5.3, especially compared to the much lower change the year before.

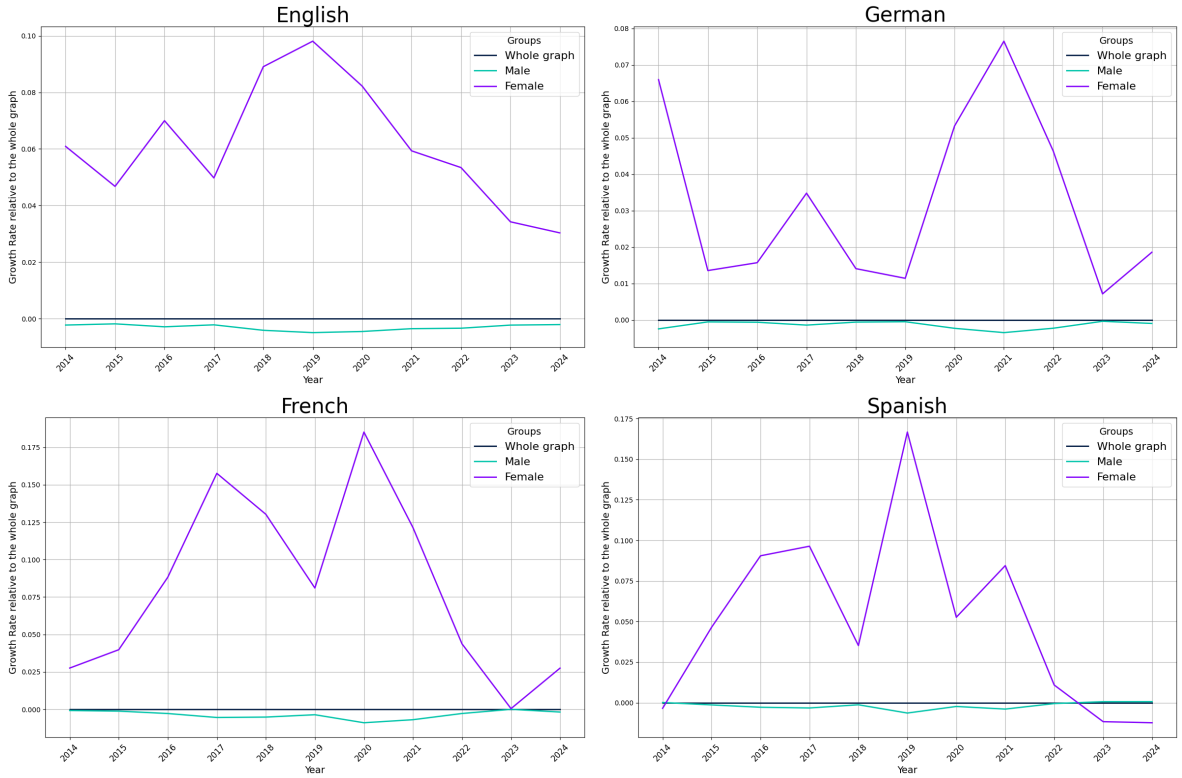


Figure 5.15: In-going edge growth rates for respectively male and female nodes for all 4 language versions from 2014-2024. Visualised as the growth rates compared to the general growth of the graph. We see that the female nodes have a larger growth rate compared to the entire network for all languages, indicating that the popularity of the female nodes is still growing faster than the popularity of all nodes.

In Figure 5.15 the in-edge growth rates are split in male and female nodes in the period 2014-2024 is seen. From the figure, we see that the female node population has a larger growth rate for in-degree compared to the entire graph. This also means that the growth rate in links to female biographies is higher than the growth rate in links to male articles.
We see that the growth of ingoing edges for women in the English graph peaks in 2019 and, after this, slowly falls but remains larger than for the whole graph. The same tendencies are seen for the other languages but with peaks in 2021 for the German, 2020 for the French and in 2019 for the Spanish. We also see that the peaks for the female in-going edge growth rates in the French and Spanish graphs are higher than for the English and German graphs. Finally, towards the end of the time period, the female growth rate is approaching the growth of entire network.

CHAPTER 6

Discussion

6.1 Data

The goal of our data scraping process was to scrape revisions for as many scientists as possible for as long a time period as possible. The ideal data would be all revisions for all scientists that has or has had a Wikipedia biography in English, German, French or Spanish in the time period 2001-2024. Why this has not been possible will be uncovered in this section.

6.1.1 Wikipedia lists and attribute detection

As mentioned in Chapter 3, our entire network is defined by Wikipedia's own Lists of scientists. These lists (all seen in Appendix A) vary from lists based on nationality such as *List of Ethiopian scientists* to more specific lists as *List of Jewish American chemists* to wider categories as the *List of Statisticians*. The lists are all created by volunteer editors and they suffer from all kinds of biases. We have, through our data collection, been aware that we most likely are missing several notable scientists in our lists, male as well as female. During our data collection phase, we searched for alternative lists to avoid adding Wikipedia's own bias to our network, but we were not able to find lists that were more comprehensive than what Wikipedia had to offer.

When detecting the gender of the article, we decided to use a limited amount of features for the classifier. This choice was made based on continuous trial and error, evaluating the precision and accuracy of the model until we landed on a satisfying result. Having learned that women made up less than one-fifth of the Wikipedia biographies, we wanted to make sure that we correctly classified as many of them as possible in our data. We did this by evaluating the model's precision and not just using the accuracy. Doing so, we experienced that when we used more word features, the precision for the female articles fell. We expect that a possible cause of this is the female articles being shorter on average than the male articles, as well as the different words used for men and women in Wikipedia as described in Chapter 2.
By limiting the features as much as we did, we were able to train a model that is very good at classifying both men and women in our dataset.

Finally, when reviewing the birth year detecting function this could be improved for better accuracy by using a classifier model as was done for gender. For our detection, we used our processed text and having removed all special characters, including hyphens, we ended up with text strings where all indications of time periods such as **1942-2001** were replaced with **19422001**. To improve the function keeping the hyphens could give us more precise information on birth years and potentially also death years. The exact birth year was, however, not important for our purposes as our main goal with including this was to indicate the time period a person lived in, which our function managed to do.

6.1.2 Word counts

Looking at the counts of female and male-related words, we found that male articles had a significantly larger percentage of male-related words than the opposite for female articles. This surprised us since male articles, on average, are longer, and we therefore expected them to include more text about other

persons, including women. It would, therefore, be interesting to investigate the relation between the article length and the percentage of ”fitting” gender words. In the same context, one could consider only using a given length of the article as input to the gender classifier, for instance, the first 1000 characters as done for the birth year function.

In context to the word counts, we found that female articles on average are shorter than the male articles. This can be a part of the explanation to why the female nodes on average have lower out-degrees than the male nodes. It can however not explain the significant lower in-degree of the female nodes.

6.1.3 Revisions

As mentioned our first goal going into the data scraping process was to scrape all revisions in the period from 2001 to today. However, this soon proved to be more time-consuming and complicated than we had anticipated. For instance, the article about Albert Einstein had 34 revisions in November 2024 alone and aimed to scrape revisions for 23 years and more than 20,000 persons. This would take a very long time to run and result in a very large data set. For context, scraping revisions for the German, French and Spanish Wikipedias took 2-3 days each and the English twice as long. This meant choosing between having many nodes on our networks but fewer revisions overall or fewer nodes and more revisions. Our priority was to scrape data for as many people as possible because we also wanted to make sure that every link we scraped would also have its own outgoing links scraped. Therefore, we chose to scrape one revision each year per article, which gave us enough data to get a sufficient overview of the development in the 23 years Wikipedia has existed.

During our data scraping process we did a number of iterations before we found the best way to scrape our data. We will not go through all of the mistakes we made in the process, but an example is one of our first attempts, where we accidentally removed names from our list of scientists when we had scraped the revisions of that person. This meant that if a person linked to another person who had already been scraped, we did not add this link to our list of links, and we gradually lost more and more information as we got through the list. This is just one example of many small mistakes in our code that resulted in several days runtime with results we could not use. Multiple times we managed to scrape all of the articles before realising mistakes in the data, and this was undeniably the biggest inconvenience with the long runtime of our revisions scraping code.
A habit that we quickly gained from this was to use the data as we scraped it. This helped us catching these kinds of mistakes early on by doing daily checks, so that we wasted a minimum amount of time and computer power. This also gave us a leg up in terms of our analysis, as we where able to start preparing some of our framework while the data was still scraping, so that it would be trialled and tested when the data scraping was finalised.

6.1.4 Non-scientist connections

For this thesis we decided to limit our network to only consist of scientists biographies. We could however have included non-biography articles to find more connections between the scientists. For instance, we imagine that using Wikipedia articles about universities of which many scientists link to from their own Wikipedia biographies, would result in serving as big hubs in our network.
We could also have extracted all links from each of person in our list and in that way expand our network to have non-biography and non-scientist articles. These networks would naturally be significantly larger than the network we ended up with and if we still only scraped the people from our list we would have to take into account that the nodes that was not from the list would not have any outgoing edges. If we had done it this way we assume that the structure of the graph would have changed and that there would be more hubs in our network.

6.1.5 Other languages

Our initial plan was to analyse the four largest four language versions which as mentioned in Section 2.1.1 is the English, German, French and Swedish if disregarding Cebuano. At the beginning of our analysis, we scraped these four Wikipedia versions for articles in our scientist list. The Swedish revisions did, however, only include 1,886 persons compared to 5,671 for the German and 4,863 for the French. Therefore we decided to consider other languages as well, and despite not being among the largest Wikipedias, Spanish is among the most spoken languages worldwide and we therefore decided to scrape the Spanish articles instead of the Swedish. This resulted in 4,484 Spanish articles, which was remarkably more than what was possible for the Swedish version.

In the parts of our analysis where the article content is used, we have only scraped the English Wikipedia articles in their current version. Looking at the article content in other languages at other times could reveal additional insights about the differences across languages and time. It would be interesting to see if the counts of female and male words are similar in other language versions or if some languages are more or less biased in that context.

Furthermore, it could be interesting to do the TF-IDF analysis on all language versions to get more local insights. There is a possibility that the German article about a German scientist might have more details than the English version, so adding this information to our data might reveal new interesting conclusions. The way we have used TF-IDF, we have used the English version for non-English networks, and though this has given us many insights on the non-English communities, the documents we analyse are more or less the same across languages when using this approach.

Another interesting aspect that we never got to investigate was the length of the articles in other languages. We were especially curious when it came to the Spanish network, where we experienced a lot fewer connections compared to the other languages. We suspect that this could be partially due to the article lengths and their content, and something that could be further researched in the future.

6.1.6 Not in data

As mentioned in Chapter 3 we go from having 28,803 unique names in our list to having revisions of 23,843 persons in the English data. Having scraped the 28,803 names from lists of Wikipedia articles, we assumed they would all, in fact, have a Wikipedia article. Some of them could be spelt differently or added to the list without having an article, but we decided not to investigate this further as we assumed that these would be some of the smaller and less-known scientists. For the other languages than English we found significantly fewer articles since many of the less famous scientists only appear in very few different languages. By using as many 'local' lists as possible, we tried to localise the less-known scientists from non-English speaking countries. This did, however, not help much as we, for instance, found the relatively long List of Swedish Scientists but still only had 1,886 persons in the Swedish network.

6.2 Network Analysis

6.2.1 Women in the network

Looking at our data, it is evident that women are under-represented in both science and Wikipedia, but looking at the subject historically this was expected. The very low percentage of women in our network could be caused both by the historically low percentage of women in science due to the discrimination they have faced in the past and are still facing today. It could also be caused by the editors since it has been found that male editors, who make up more than 90% of all editors, are more likely to write about men than women. In our analysis, we found that articles about female scientists that do exist

are added later and born later than the articles about male scientists. This indicates that the overall amount of female scientists could be growing and thereby also the representation of female scientists in places like Wikipedia.

6.2.2 Communities

For this thesis, we chose to divide the network into communities using InfoMap. In the process of doing so, we also considered using the method Louvain for community detection. However, InfoMap turned out to give us fewer, larger communities, which was more what we were looking for. We found that the two results of the two methods were quite similar in content, indicating that the members of our communities also have something in common according to the Louvain method. In this thesis, we did not focus on the community detection methods as much as the structures and content of the detected communities, but it would be interesting to combine the two methods and look further into what extent they overlap. Furthermore, none of these methods allows for members to be part of more than one community, while in the case of our network, a person could be relevant in more than one scientific field and, therefore, also in more than one community.

6.2.3 TF-IDF and field of study

Having created our communities, we did a TF-IDF analysis of the content of the articles in each community. This resulted in a detection of fields for each of our 5 focus communities. We have not investigated the correlation between the results from our TF-IDF analysis and the lists our scientists came from in the first place, but the list origin of each scientist could have been added to our data to receive even more indication of the field of study of each scientist in our network.

6.2.4 Network structure and robustness

As we moved on in our analysis and looked at the structure of all the graphs, we found that across all languages, the female nodes were much worse connected than the men. The percentage of female nodes removed in each of the k cores revealed that they were not as crucial in the hubs as the male nodes and, therefore, that the male nodes are crucial for the network robustness. We also found that for the three other languages that have a similar size, there was a great difference in connectivity and robustness. We found that the Spanish network was much less connected than the German and French networks, which showed to have stronger connected cores and more highly connected hubs. This surprised us initially, and because of this, we can't lump them all together and conclude that the smaller-sized networks are less connected and robust than the big English network.

6.2.5 Bursty growth

We found that when completing the permutation test, all four networks exhibited bursty growth over the years. This was not entirely surprising to us as when we investigated the nodes' individual out-degree over time, we found that most of the nodes grew out-edges in a more rapid pattern than evenly over time. Considering the assumptions about the expected growth likely growing evenly in the graph due to the probability of a node receiving an out-edge from Equation 4.6.1, the result of this investigation already suggested that our network might deviate from the null model.

When investigating the growth of the individual nodes over time, we also found that the out-degree growth of a node could be vastly different across language versions. This did surprise us, and we believe that this could be due to the biographies in English not just being translated but actually written in the other language.

As mentioned earlier, we chose to scrape one revision per year for each scientist in order to have as many nodes in our network as possible. For our growth analysis and detection of bursty growth this meant that detecting patterns happening within the years was impossible with our data. As we found, 27-38% of the nodes that had links added in the English graph had more than one link added during the year. And this percentage is almost twice as big for the other languages. If the links that were added at the same year, were added at the same time or right after each other, they would be considered bursty growth. However, having multiple timestamps throughout the year for our revisions might not entirely solve our problem, since we still wouldn't be sure that growth wouldn't happen at the same time in the network even with more timestamps.

6.2.6 Growth rates

When comparing the growth rates for each of the focus communities with the remaining network, we found that the node and edge growth rates for the English communities over time seems to fluctuate less over the years and center closely around the entire networks' growth rates. We suspect that as the network has grown larger over the years, there is no longer targetted bursts of interest in these communities, indicating that they have reached a high level of maturity. We also found that the Mathematics community, that has an overall higher growth rate compared to the growth rate of the rest of the graph, also had the members with the youngest age and smaller hubs. This could be a reason why it does not seem to mature as fast as the other communities and we suspect that because our List of Mathematicians was much more detailed than the other Lists of scientists, this community includes more people of lower notability than the other communities.
We experienced that for all networks except the Spanish, the mathematics communities had higher growth rates compared to the rest of the networks. This aligns with what we expected would happen with the English network, since the list we used to scrape Mathematicians was by far the most detailed. However, since we were not able to find all scientists across the other language versions of Wikipedia and we know that the German and French mathematics communities were not as big as the English one, we were surprised that they still experienced higher growth rates. While we found that the individual nodes does not necessarily grow the same way across all languages, the fact that the mathematics community continues to grow with a higher growth rate in all three networks suggests that there is similarities in how communities grow.

The in-going edge growth rate split on gender revealed that growth rates for the female nodes in the period 2014-2024 were larger than the growth rate of the entire graph. We know from chapter 2 that there have been initiatives to fight gender bias on Wikipedia and this could be a reason for the bigger growth rate. But the number of female biographies compared to the number of biographies in the entire graph are also so much smaller, that big changes will seem even bigger because it was uneven to begin with.

CHAPTER 7

Conclusion

The overall scope of this thesis was to shed light on how Wikipedia networks of scientists differed across languages, how these networks grew, and if this is different for female scientists than male scientists. To do so, we scraped revision data to construct networks and analysed them in terms of structure and growth.

The data scraping process proved to be an ongoing challenge. We did several iterations before having a working function to scrape the needed revisions data. The final result of this process was an extensive data set with many successfully scraped articles, but unfortunately, also with flaws, as explained in the discussion. Classifying the gender of the scientists was crucial to our research, and to get as precise classifications as possible, we manually classified 600 articles and trained our model on these articles. With an accuracy of 97.32%, we are pleased with the result and expect that only a few articles in our data have been misclassified.
Going into the analysis, we created networks using the hyperlink structure for each language and each year in 2001-2024, giving a total of $24 \times 4 = 96$ network snapshots. The networks for the year 2024 were the basis for the final graph which we used for the initial network analysis. The snapshots were then used to provide a timeline for the networks, where we could analyse the structure and growth by finding the difference between them every year.

Looking at the structure of the largest network, which is the English network in 2024, it is clear that the female members of our graph are significantly less connected to other nodes than the male members. This is clear from the k-core analysis, where the majority of the female nodes were revealed to lie in the periphery, while highly connected male nodes were the most crucial to the robustness. This is also clear from the degree distribution where we saw how female nodes had both smaller in and out degrees than the male nodes. This picture continues when analysing the other languages where the story seems similar across versions. However, despite being much smaller than the English network, we found that the German graph was well-connected compared to the Spanish and French networks of similar size. The Spanish network differed by being much less connected than the other graphs, meaning that the number of links per article was smaller, suggesting that the network was built by shorter articles with fewer details. Furthermore, it was possible to group the nodes into communities, where we found topics using TF-IDF, confirming that the members had fields of study and article content in common. We found that much of the same community structure reoccurred across the languages and the overall field of each of our communities matched.

Looking at the growth of the networks, we were able to detect bursty growth patterns in all four networks. This result tells us that the networks do not grow new edges in random places but that there exists a pattern where new edges grow where edges recently have grown. Furthermore, we saw that the communities grew at different speeds over time, indicating that different areas of the network experienced more interest at times compared to the entire network.

In this thesis, we have found several aspects that indicate that the female articles are worse connected and have less significance in the network. The lack of female presence in our network is, however, not only a product of Wikipedia's bias but also a field that historically has been dominated by men. It is difficult to determine what is due to gender bias and what is due to history, but it is clear that women

are under-represented in many aspects of the network of Wikipedia scientists.

APPENDIX A

List of scraped lists

African educators	List of Mathematicians(A)	Nigerian scientists and scholars
Argentines*	List of Mathematicians(B)	Ornithologists
Armenian scientists	List of Mathematicians(C)	Pakistani scientists
Atheists in science and technology	List of Mathematicians(D)	Paleontologists
Austrian scientists	List of Mathematicians(E)	Pathologists
Authors published under the ICZN	List of Mathematicians(F)	People by Erdos number
Azerbaijani scientists and philosophers	List of Mathematicians(G)	Pharmacists
Bangladeshi scientists	List of Mathematicians(H)	Photochemists
Biologists	List of Mathematicians(I)	Physicists
Biophysicists	List of Mathematicians(J)	Pre-modern Arab scientists and scholars
Brazilian scientists	List of Mathematicians(K)	Pre-modern Iranian scientists and scholars
British Jewish scientists	List of Mathematicians(L)	Presidents of the Geological Society of London
Byzantine scholars	List of Mathematicians(M)	Psephologists
Chemists	List of Mathematicians(N)	Quakers in science
Chinese scientists	List of Mathematicians(O)	Quantum gravity researchers
Christians in science and technology	List of Mathematicians(P)	Researchers at Racah Institute
Climate scientists	List of Mathematicians(Q)	Rheologists
Cognitive scientists	List of Mathematicians(R)	RNA Tie Club
Computer scientists	List of Mathematicians(S)	Romanians*
Cornish scientists	List of Mathematicians(T)	Runologists
Cosmologists	List of Mathematicians(U)	Russian scientists
Czechs*	List of Mathematicians(V)	Scientists in medieval Islamic world
Ecologists	List of Mathematicians(W)	Soil scientists
Egyptian scientists	List of Mathematicians(X)	Spectroscopists
Estonian scientists	List of Mathematicians(Y)	Statisticians
Ethiopian scientists	List of Mathematicians(Z)	Swedish scientists
Female scientists	Medieval and pre-modern Persian doctors	Systems scientists
French scientists	Meteorologists	Taxonomic authorities by name
Geologists	Mineralogists	Undersea explorers
Geophysicists	National Medal of Science laureates	Welsh scientists
Indian scientists	Nepalese scientists	Women in chemistry
Italian scientists	Neurochemists	Women in mathematics
List of Jewish American chemists	Neurologists and neurosurgeons	Women in statistics
Loop quantum gravity researchers	New Zealand scientists	

Table A.1: Links to all lists used to build the dataset. *Only sections about scientists of the given nationality are included

APPENDIX B

Birth years

Community	25th Percentile	Median Birth Year	75th Percentile
1	1897	1920	1941
2	1927	1945	1966
3	1786	1822	1858
4	1280	1599	1759
5	1865	1904	1931
6	1902	1929	1948
7	1940	1954	1974
8	1896	1926	1950
9	1857	1886	1908
10	1889	1919	1945
11	1885	1907	1936
12	1933	1945	1960
13	1919	1939	1956
16	1887	1913	1944

Table B.1: Estimated birth years of communities larger than 50 members

APPENDIX C

TF-IDF: Top Words

English

Community	Size	% female	Top 30 words
1	2517	12.59	quantum, nuclear, atom, that, hi, particl, she, energi, laboratori, had her, it, theoret, physicist, physic, would, be, thi, space, mechan institut, electron, develop, chemic, soviet, award, scientif, war, cambridg, experi
10	280	16.43	dinosaur, fossil, paleontolog, bird, expedit, museum, vertebr, she, her, anim dive, specimen, that, speci, beeb, evolut, natur, gorilla, excav, geolog ocean, skull, hi, had, california, milankovi, chimpanze, bone, collect, human
11	128	15.63	ramanujan, mathemat, cambridg, edinburgh, notebook, function, number, she, whittak, triniti her, wrangler, hi, displaystyl, form, dunde, tripo, that, conjectur, isbn modular, algebra, fellow, india, 1, milnethomson, had, british, it, architectur
12	109	13.76	logic, cardin, comput, algebra, mathemat, set, vopnka, ramrezgarofalo, displaystyl, isbn politi, czech, integr, axiom, hedrln, babuka, prove, berkeley, conjectur, cambridg pragu, infinit, proof, nonstandard, topolog, finit, algorithm, symbol, vocera, forc
13	157	15.29	chemic, santosdumont, reaction, synthesi, molecul, structur, protein, molecular, award, enzym chemist, acid, that, dna, organ, hi, she, compound, harvard, nmr metal, use, patent, laboratori, energi, complex, organometal, metathesi, catalysi, california
16	58	12.07	bird, bombay, india, ornitholog, bnh, specimen, speci, natur, calcutta, ornithologist meinertzhagen, museum, collect, conserv, ibi, bengal, geograph, jbnh, expedit, wildlif british, nicobar, zoolog, gandhi, lanka, that, sri, hi, birdwatch, tropic
2	2820	20.00	mathemat, algebra, geometri, topolog, differenti, conjectur, she, manifold, displaystyl, equat mr, space, function, math, group, princeton, problem, isbn, her, quantum hi, comput, vol, moscow, geometr, proof, prove, number, pp, s2cid
3	2097	8.63	hi, speci, collect, that, her, had, museum, bird, natur, plant it, him, she, expedit, specimen, but, thi, be, botan, naturalist fossil, und, geolog, which, medic, british, would, insect, zoolog, or
4	1528	6.68	that, astronom, translat, arab, it, hi, greek, astronomi, latin, philosoph hadith, him, god, be, treatis, thi, have, or, text, commentari but, had, write, are, caliph, all, would, observ, which, astrolog
5	1302	16.74	algebra, geometri, mathemat, she, function, gttingen, displaystyl, und, her, number hi, mr, group, math, ber, isbn, chess, that, mathematik, vol problem, proof, differenti, poincar, mactutor, equat, finit, mathematisch, graph, logic
6	1292	12.69	logic, that, econom, comput, it, hi, cognit, be, languag, film had, philosoph, human, mind, her, thi, she, would, isbn, idea are, but, can, or, intellig, what, argu, concept, have, they
7	1015	17.73	comput, graph, algorithm, mathemat, acm, erd, she, combinator, problem, number program, displaystyl, conjectur, geometri, complex, languag, cryptographi, proof, game, her award, hi, combinatori, random, algebra, ha, isbn, design, softwar, ieee
8	710	20.70	genet, evolut, statist, that, bird, she, evolutionari, isbn, her, natur hi, human, bbc, it, be, anim, speci, popul, had, harvard zoolog, ornitholog, but, cambridg, would, thi, british, have, ha, about
9	175	9.71	zbl, matematica, nazional, dei, geometri, accademia, mr, lincei, scienz, rendiconti function, matematich, seri, algebra, pp, pisa, differenti, bologna, mathemat, vol equat, levicivita, roma, palermo, atti, matematico, funzioni, sulla, italiana, teoria

Table C.1: Top 30 TF-IDF words for English communities larger than 50 members

German

Community	Size	% female	Top 30 words
1	1943	10.55	that, her, she, quantum, logic, hi, it, nuclear, econom, mathemat be, atom, isbn, had, algebra, wa, would, thi, but, particl geometri, which, war, cambridg, philosoph, energi, they, und, is, displaystyl
10	78	10.26	rocket, milankovi, ocean, geolog, geophys, space, kppen, continent, braun, underwat film, drift, dive, expedit, that, sea, earth, geologist, she, song hi, graphen, cameron, tecton, usg, lehrer, titan, peenemnd, thobald, oceanograph
11	80	6.25	quantum, martinlf, mathemat, inflat, algebra, grtner, logic, statist, random, atom inflationari, princeton, energi, mechan, phi, oper, stochast, math, rrdam, cohentannoudji calgebra, heghkrohn, univers, laser, youtub, sundberg, probabl, strmer, spectral, isbn
13	73	12.33	comput, darmstadt, algorithm, hndler, siam, program, linear, she, algebra, und ieee, enigma, optim, languag, mathemat, acm, method, algol, fr, woniakowski glacier, highperform, softwar, mathematik, matrix, appli, cipher, althfer, machin, equat
14	93	5.38	atmospher, meteorolog, global, chang, weather, ipcc, temperatur, rockstrm, carbon, geophys dioxid, that, ocean, isotop, ozon, co2, bbc, forecast, ice, environment planet, cosmic, circul, satellit, scientif, model, solar, climat, greenhous, predict
15	58	17.24	blake, mead, samoa, uexkll, samoan, anthropolog, cognit, human, popul, her that, pp, she, cultur, sexual, languag, borlaug, mind, anthropologist, wheat wavelet, argu, coevolut, freeman, no, mosquito, food, pinker, pask, commun
2	1066	12.38	algebra, geometri, mathemat, conjectur, topolog, mr, differenti, group, she, manifold space, math, poincar, vol, cohomolog, displaystyl, function, isbn, proof, annal princeton, represent, equat, her, mathmatiqu, pp, problem, geometr, hi, curv
3	1491	6.30	speci, bird, natur, museum, that, hi, her, collect, she, plant zoolog, specimen, expedit, had, naturalist, it, ornitholog, anim, fossil, wa be, evolut, und, him, botan, but, thi, british, vol, geolog
4	813	8.24	that, her, astronom, hi, it, latin, astronomi, had, she, translat but, be, him, thi, wa, treatis, observ, would, motion, philosoph which, displaystyl, mathemat, edit, have, or, use, all, vol, wrote
5	460	10.87	mathemat, graph, erd, combinator, conjectur, number, comput, autism, she, problem equat, differenti, algorithm, geometri, algebra, function, hungarian, mr, baroncohen, proof displaystyl, her, isbn, math, probabl, combinatori, princeton, prove, finit, hi
6	148	10.14	manifold, curvatur, geometri, quantum, metric, riemannian, differenti, conjectur, flow, particl khler, mathemat, pakistan, symmetri, theoret, princeton, topolog, space, yau, graviti zbl, khlereinstein, scalar, geometr, dimens, proof, theori, mr, gaug, problem
7	197	6.60	comput, jackson, program, that, languag, softwar, intellig, album, logic, it brown, hi, mit, oper, ai, artifici, piaget, be, perform, develop guitar, unix, gnu, use, would, cognit, had, isbn, learn, machin
8	189	12.17	comput, algorithm, googl, acm, wireless, patent, electr, that, displaystyl, complex cryptographi, ai, machin, secur, tesla, hi, microsoft, radio, telephon, amazon learn, she, quantum, it, billion, marconi, network, program, telegraph, award
9	112	6.25	dinosaur, fossil, expedit, paleontolog, beeb, vertebr, specimen, museum, skull, speci osmlska, cope, andrewsi, nordenskild, bone, kansa, paleontologist, cretac, theropod, frill anim, that, bird, kielanjaworowska, china, reptil, geolog, cyclotron, velociraptor, djadokhta

Table C.2: Top 30 TF-IDF words for German communities larger than 50 members

French

Community	Size	% female	Top 30 words
1	1229	13.34	logic, econom, mathemat, that, she, her, algebra, it, be, geometri languag, comput, philosoph, isbn, function, polit, idea, hi, argu, problem object, mind, set, displaystyl, can, are, thi, human, had, cognit
10	163	11.04	mathemat, ramanujan, number, displaystyl, function, dymaxion, geometri, conjectur, cambridg, problem she, analyt, integ, isbn, algorithm, comput, space, algebra, dome, sum prove, notebook, frhlich, her, award, siev, geodes, polynomi, fuller, hi
11	121	10.74	meteorolog, weather, expedit, beeb, ocean, atmospher, nordenskild, her, geophys, she sea, forecast, oceanographi, potter, permafrost, ice, film, geolog, antarct, carson global, environment, it, pheasant, that, acoust, tale, predict, pesticid, silent
12	68	4.41	china, shen, armillari, clock, astronom, sphere, calendar, zhang, japanes, oclc needham, magnet, celesti, sanp, it, bc, astronomi, ancient, treatis, waterpow that, clepsydra, water, garden, ricci, song, mathemat, calcul, write, mechan
2	854	11.48	quantum, atom, nuclear, she, particl, her, that, theoret, it, energi mechan, physicist, laboratori, had, dna, rel, bomb, space, physic, would be, hi, neutron, electron, radiat, isbn, thi, war, mathemat, comput
3	1343	6.85	speci, bird, natur, museum, that, collect, her, plant, it, anim she, hi, had, zoolog, specimen, evolut, fossil, expedit, naturalist, him human, british, botan, be, insect, ornitholog, thi, genet, but, scientif
4	959	18.67	mathemat, geometri, algebra, differenti, topolog, conjectur, manifold, she, equat, mr math, group, function, space, princeton, geometr, isbn, problem, cohomolog, curvatur proof, her, comput, represent, prove, curv, symplect, surfac, partial, riemannian
5	712	8.85	astronom, it, that, her, astronomi, translat, she, hi, latin, motion him, star, be, had, treatis, observ, thi, film, but, would philosoph, arab, have, god, calcul, or, mathemat, write, appear, which
6	246	16.26	graph, erd, mathemat, combinator, comput, geometri, problem, she, conjectur, hungarian game, mr, function, algorithm, algebra, number, combinatori, proof, isbn, displaystyl math, her, hebrew, budapest, cambridg, topolog, space, prove, discret, set
7	136	13.97	autism, electr, that, she, her, radio, it, had, wireless, autist would, patent, hi, baroncohen, light, war, be, wave, milankovi, show game, him, gibbss, but, isbn, malaria, roosevelt, thi, us, asperg
8	85	24.71	barabsi, graph, wavelet, she, mathemat, caldern, network, lovsz, equat, optim bueno, her, air, grtschel, comput, tao, problem, cand, differenti, argentina award, conjectur, insulin, enzym, function, algebra, cliqu, nonlinear, organel, albertlszl
9	166	16.87	comput, algorithm, acm, cryptographi, complex, program, languag, unix, quantum, she graph, displaystyl, award, ieee, mathemat, secur, cryptograph, problem, proof, theoret softwar, scientist, code, mit, berkeley, game, npcomplet, machineri, random, symposium
8	189	12.17	comput, algorithm, googl, acm, wireless, patent, electr, that, displaystyl, complex cryptographi, ai, machin, secur, tesla, hi, microsoft, radio, telephon, amazon learn, she, quantum, it, billion, marconi, network, program, telegraph, award
9	112	6.25	dinosaur, fossil, expedit, paleontolog, beeb, vertebr, specimen, museum, skull, speci osmlska, cope, andrewsi, nordenskild, bone, kansa, paleontologist, cretac, theropod, frill anim, that, bird, kielanjaworowska, china, reptil, geolog, cyclotron, velociraptor, djadokhta

Table C.3: Top 30 TF-IDF words for French communities larger than 50 members

Spanish

Community	Size	% female	Top 30 words
1	603	6.97	nuclear, atom, quantum, energi, particl, laboratori, radioact, bomb, uranium, had war, that, she, it, neutron, physicist, her, rel, radiat, mechan would, physic, chemic, electron, reactor, experi, be, element, theoret, wave
10	119	14.29	comput, dodgson, bernersle, program, web, nabokov, game, she, cryptographi, algorithm tile, mathemat, acm, network, corbat, award, concurr, softwar, mit, chess column, complex, it, gardner, her, languag, secur, problem, oper, pentagon
11	96	19.79	medic, diseas, lister, her, patholog, hospit, she, blood, vaccin, cell tissu, jexblak, nurs, antisept, surgeri, patient, wound, clinic, pruvotfol, tuberculosi pasteur, acid, surgeon, caus, infect, experi, that, physiolog, edinburgh, had
12	62	6.45	cognit, piaget, comput, facebook, genet, psycholog, intellig, googl, child, develop statist, learn, mental, she, experiment, popul, that, program, development, fisher psychologist, human, disson, concept, network, eugen, it, design, knowledg, be
13	70	10.00	touchdown, colt, comput, nfl, yard, game, quarterback, bronco, show, pass bowl, unix, afc, network, letterman, krusti, algorithm, episod, ai, intercept acm, grammer, program, week, super, win, threw, player, man, simpson
18	68	8.82	geometri, manifold, curvatur, conjectur, perelman, differenti, riemannian, topolog, algebra, yau mathemat, 3manifold, geometr, hypersurfac, mr, flow, surfac, proof, metric, space zbl, shiingshen, minim, dimens, hyperbol, mongeampr, scalar, mathematica, bundl, math
2	1175	10.04	logic, econom, mathemat, algebra, geometri, she, function, displaystyl, her, isbn it, that, problem, differenti, comput, mr, philosoph, be, set, number space, equat, war, had, georgescuroegen, can, vol, hi, concept, cambridg
3	687	4.80	astronom, it, astronomi, that, latin, had, motion, god, hi, be observ, him, translat, her, treatis, chtelet, but, would, philosoph, displaystyl thi, have, calcul, newton, letter, star, telescop, arab, light, or
4	725	10.62	speci, her, plant, she, museum, collect, natur, fossil, bird, anim specimen, had, zoolog, expedit, it, botan, naturalist, garden, british, that hi, botanist, geolog, darwin, him, botani, travel, be, would, but
5	276	10.87	dna, molecular, genet, chemic, her, molecul, she, structur, xray, laboratori that, had, protein, acid, discoveri, medic, award, physiolog, govern, reaction genom, enzym, rocket, it, helix, cell, lab, would, war, develop
6	148	4.73	chemic, heat, humboldt, edinburgh, atom, plant, goeth, whler, electr, element her, kekul, und, liebig, magnet, dalberti, carnot, compound, had, chemist it, hi, she, that, temperatur, light, physiolog, experi, xray, him
7	151	13.91	evolut, evolutionari, genet, human, bird, that, she, isbn, cognit, natur her, languag, linguist, pakistan, god, teilhard, speci, it, intellig, behavior harvard, critic, religion, ha, be, view, argu, anim, particl, philosoph
8	143	13.99	quantum, graviti, cosmolog, gravit, theoret, mathemat, hole, univers, topolog, rel particl, theori, space, physicist, mechan, manifold, princeton, gaug, isbn, ligo cambridg, geometri, she, equat, algebra, ktheori, spacetim, physic, loop, symmetri
9	130	12.31	freud, her, psychoanalysi, psychoanalyt, human, it, sexual, that, phenomenolog, object be, she, merleauponti, heidegg, gestalt, psycholog, idea, lacan, cultur, isbn husserl, therapi, experi, leonardo, individu, anthropolog, concept, cognit, psychotherapi, lvistrauss

Table C.4: Top 30 TF-IDF words for Spanish communities larger than 50 members

Bibliography

[1] Similarweb. August 2024. URL: https://www.similarweb.com/top-websites/ (visited on 12 August 2024) (cited on pages 1, 3).

[2] WikiMedia. *Editor Survey Report - April 2011*. URL: https://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report_-_April_2011.pdf (cited on pages 1, 5).

[3] Humaniki. URL: https://humaniki.wmcloud.org/search (cited on pages 1, 3).

[4] Claudia Wagner et al. "It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia." In: *The International AAAI Conference on Web and Social Media (ICWSM2015), Oxford* (May 2015) (cited on pages 1, 4, 6, 11).

[5] WHO. *Gender and health*. URL: https://www.who.int/health-topics/gender#tab=tab_1 (visited on 13 August 2024) (cited on page 1).

[6] Maximilian Klein Piotr Konieczny. "Gender gap through time and space: A journey through Wikipedia biographies via the Wikidata Human Gender Indicator." In: *New Media Society, 20(12)* (2018) (cited on page 1).

[7] Mahzarin R. Banaji Tessa E.S. Charlesworth. "Gender in Science, Technology, Engineering, and Mathematics: Issues, Causes, Solutions." In: *The Journal of Neuroscience* (September 2019) (cited on page 2).

[8] Cary Funk and Kim Parker. *Women and Men in STEM Often at Odds Over Workplace Equity*. 2018. URL: https://www.pewresearch.org/social-trends/2018/01/09/blacks-in-stem-jobs-are-especially-concerned-about-diversity-and-discrimination-in-the-workplace/ (visited on 20 August 2024) (cited on page 2).

[9] Arya Min and Jessica Ventre. *The Past, Present and Future of Women in STEM*. URL: https://edventures.com/blogs/stempower/the-past-present-and-future-of-women-in-stem (visited on 17 December 2024) (cited on page 2).

[10] Anthony Martinez and Cheridan Christnacht. *Women Making Gains in STEM Occupations but Still Underrepresented*. URL: https://www.census.gov/library/stories/2021/01/women-making-gains-in-stem-occupations-but-still-underrepresented.html (visited on 17 December 2024) (cited on page 2).

[11] UNESCO. *The gender gap in science: status and trends, February 2024*. URL: https://unesdoc.unesco.org/ark:/48223/pf0000388805 (visited on 17 December 2024) (cited on page 2).

[12] World Bank Group. *Labor force*. URL: https://data.worldbank.org/indicator/SL.TLF.TOTL.FE.ZS (visited on 17 December 2024) (cited on page 2).

[13] The Editors of Encyclopaedia Britannica. *Wikipedia*. August 2024. URL: https://www.britannica.com/topic/Wikipedia (visited on 12 August 2024) (cited on page 3).

[14] Haifeng Zhang Yuqing Ren and Robert E. Kraut. "How Did They Build the Free Encyclopedia? A Literature Review of Collaboration and Coordination among Wikipedia Editors." In: *ACM Trans. Comput.-Hum. Interact. 31, 1, Article 7* (November 2023) (cited on page 3).

[15] Wikipedia. *Cebuano Wikipedia*. URL: https://en.wikipedia.org/wiki/Cebuano_Wikipedia (visited on 26 December 2024) (cited on page 3).

[16] Wikipedia. *List of Wikipedias*. URL: https://en.wikipedia.org/wiki/List_of_Wikipedias (visited on 26 December 2024) (cited on page 3).

[17] Statista. *Worldwide visits to Wikipedia.org from July to December 2023*. 2024. URL: https://www.statista.com/statistics/1259907/wikipedia-website-traffic/ (visited on 12 August 2024) (cited on page 3).

[18] David Garcia Claudia Wagner Eduardo Graells-Garrido and Filippo Menczer. "Women through the glass ceiling: gender asymmetries in Wikipedia." In: *EPJ Data Science* (2016) (cited on pages 3–5).

[19] Hannah Brückner Julia Adams and Cambria Naslund. "Who Counts as a Notable Sociologist on Wikipedia? Gender, Race, and the "Professor Test"." In: *Socius, 5* (2019) (cited on page 4).

[20] Francesca Tripodi. "Ms. Categorized: Gender, notability, and inequality on Wikipedia." In: *Sage Journals Home Volume 25, Issue 7* (2021) (cited on pages 4, 6).

[21] Wikipedia. *Wikipedia:Too soon*. URL: https://en.wikipedia.org/wiki/Wikipedia:Too_soon (visited on 20 August 2024) (cited on page 4).

[22] Rebecca Zhang Mackenzie Emily Lemieux and Francesca Tripodi. ""Too Soon" to count? How gender and race cloud notability considerations on Wikipedia." In: *Big Data Society, 10(1)* (2023) (cited on page 4).

[23] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. *First Women, Second Sex: Gender Bias in Wikipedia*. New York, NY, USA, 2015. URL: https://doi.org/10.1145/2700171.2791036 (cited on page 4).

[24] Wikipedia. *Help:Editing*. URL: https://en.wikipedia.org/wiki/Help:Editing (visited on 20 October 2024) (cited on page 4).

[25] WikiMedia. URL: https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/meta.tab (cited on page 5).

[26] Y-H Eom et al. "Interactions of Cultures and Top People of Wikipedia from Ranking of 24 Language Editions." In: *PLoS ONE* 10.3 (2015) (cited on page 5).

[27] Nicole Torres. "Why Do So Few Women Edit Wikipedia?" In: *Harvard Business Review* (2016) (cited on pages 5, 6).

[28] WikiMedia. *Community Insights/2018 Report/Contributors* (cited on pages 5, 6).

[29] Shyong (Tony) K. Lam et al. "WP:clubhouse? an exploration of Wikipedia's gender imbalance." In: *Association for Computing Machinery* (2011). URL: https://doi.org/10.1145/2038558.2038560 (cited on page 5).

[30] The Guardian. *The Guardian view on Wikipedia: evolving truth*. URL: https://www.theguardian.com/commentisfree/2014/aug/07/guardian-view-wikipedia-evolving-truth (visited on 16 December 2024) (cited on page 5).

[31] Wikipedia. URL: https://en.wikipedia.org/wiki/Edit-a-thon (cited on page 6).

[32] 500 women scientists. *Wikipedia Edit-a-Thons*. URL: https://500womenscientists.org/wikipedia-editathon (visited on 3 November 2024) (cited on page 6).

[33] Wiki Education. *500 women Scientists Wikithons*. URL: https://outreachdashboard.wmflabs.org/campaigns/500_women_scientists_wikithons/programs (visited on 3 November 2024) (cited on page 6).

[34] Wikipedia editors. *List of mathematicians (A)*. URL: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Mathematics/List_of_mathematicians_(A) (visited on 5 September 2024) (cited on page 8).

[35] Wikipedia editors. *List of scientists*. URL: https://en.wikipedia.org/wiki/Lists_of_scientists (visited on 5 September 2024) (cited on page 8).

[36] Arthur V. Ratz. "Multinomial Na ve Bayes' For Documents Classification and Natural Language Processing (NLP)." In: *Towards Data Science* (May 2021). URL: https://towardsdatascience.com/multinomial-na%C3%AFve-bayes-for-documents-classification-and-natural-language-processing-nlp-e08cc848ce6 (cited on page 10).

[37] Scikit Learn. *Sklearn (Documentation).* URL: https://scikit-learn.org/1.5/api/sklearn.html#module-sklearn (visited on 16 December 2024) (cited on page 10).

[38] Per B. Brockhoff et al. *Introduction to Statistics at DTU.* 2018 (cited on page 14).

[39] Albert Laszlo Barabasi. *Network Science.* 2016. URL: https://networksciencebook.com/ (visited on 1 August 2024) (cited on pages 17, 20–22).

[40] M. et al Bell. "Network growth models: A behavioural basis for attachment proportional to fitness." In: *Nature* (2017) (cited on page 17).

[41] Pieter J. Swart Aric A. Hagberg Daniel A. Schult. *Exploring network structure, dynamics, and function using NetworkX.* URL: https://networkx.org/documentation/latest/index.html (visited on 3 October 2024) (cited on page 17).

[42] Gephi. *Gephi Tutorial Layouts.* URL: https://gephi.org/tutorials/gephi-tutorial-layouts.pdf (visited on 12 December 2024) (cited on page 18).

[43] Rahmat Ullah Orakzai. *What Is the K-Core of a Graph?* May 2023. URL: https://www.baeldung.com/cs/graph-k-core? (visited on 13 December 2024) (cited on page 21).

[44] Anton Holmgren Daniel Edler and Martin Rosvall. *infomap module.* 2020. URL: https://mapequation.github.io/infomap/python/ (visited on 13 November 2024) (cited on page 22).

[45] Geeks for Geeks. *Understanding TF-IDF (Term Frequency-Inverse Document Frequency).* URL: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/ (visited on 15 October 2024) (cited on page 23).

[46] Yan Holtz. *Sankey Diagram.* URL: https://www.data-to-viz.com/graph/sankey.html (visited on 23 December 2024) (cited on page 23).

[47] Plotly. *Parallel Categories Diagram in Python.* URL: https://plotly.com/python/parallel-categories-diagram/ (visited on 23 December 2024) (cited on page 23).

[48] Geeks for Geeks. *Understanding the Confusion Matrix in Machine Learning.* URL: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/ (visited on 23 December 2024) (cited on page 23).

[49] Michael Berk. *How to use Permutation Tests.* URL: https://towardsdatascience.com/how-to-use-permutation-tests-bacc79f45749 (visited on 18 December 2024) (cited on page 33).

[50] Jonas Lybker Juul. *Bursty Growth.* Personal communication. 2022 (cited on page 33).