Gender Bias on Wikipedia

An analysis of the affiliation network

Feli Nicolaes
10542442

Bachelor thesis
Credits: 18 EC

Bachelor Opleiding Kunstmatige Intelligentie

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Dr. M. J. Marx

ILPS, IvI
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

2016-06-24

Contents

Abstract

Wikipedia is a popular website where users can write biographies about famous people. While it is read by an almost equal number of men and women, previous research found two gender gaps: one in the users and one in the person pages. Other research has also shown that the edit activity is different for both genders.

Some suggested homophily was present between users and person pages: females would mostly edit female person pages and the other way around.

This thesis has combined several datasets to find out more about the editing behaviour of both genders by not looking at the users and pages separately, but looking at the affiliation network between them. It was found that while a clear gender gap is present in the users and person pages, no evidence of the previously suggested homophily was found.

No homophily was found in the affiliation network, both when looking at the amount of edits made or the amount of sessions spent on a page. This suggests the two gender gaps might not be as related as some expected.

1 Introduction

Wikipedia calls itself the 'Free Encyclopedia', which means that anyone who can access the site, can edit it. This gives everyone the possibility to add information to already existing subjects or to add a new subject. It is ranked among the ten most popular websites, making it one of the largest online open-source, community-driven projects of all time [Hill and Shaw, 2013].

Apart from the fact many people use Wikipedia, Wikipedia also forms the foundation of many knowledge bases such as YAGO, DBpedia, Google's Knowledge Graph and IBM Watson, which is why it is important to know whether Wikipedia is an unbiased knowledge bank.

When looking at Wikipedia, two gender gaps can clearly been seen. There is a gender gap in person pages and a gender gap in users, both of which will be explained in the following paragraphs.

Wikipedia pages can be classified into different types, one of which is person pages, these are pages discussing non-fictional people. This includes a wide variety of people, from politicians to ice skaters. When looking at these person pages, a large gender gap can be detected: only 15% of Wikipedia's person pages are about women [Graells-Garrido et al., 2015].

Wikipedia is edited by its users, commonly referred to as editors or Wikipedians. Previous research has shown a gender disparity is also present within these users: only 13-16% of the Wikipedia editors is female and they make up an even smaller percentage of all edits: only 9% of all edits were made by a female user [Lam et al., 2011, Antin et al., 2011]. When only looking at the bottom 75% of editors by activity level, men and women make a similar number of revisions, but when looking at the most active editors, the gender gap becomes even more pronounced. This shows that of all female editors, only few are avid editors.

Jimmy Wales, the founder of Wikipedia, has previously said this gender gap in users is a problem: "The main thing is to bring in people of all different backgrounds. If you do that, you increase the knowledge base of the site, which can only be a good thing. At the moment, we are relatively poor in a few areas; for example, biographies of famous women through history and issues surrounding early childcare".[1] Wales thus suggests homophily is present in this network, which would mean women mostly edit pages about women. This also assumes that if the number of female editors were to increase, so would the number of female person pages.

While both the gender gap in the person pages and the gender gap in the editor community have often been researched, no attempt has been made to find the correlation between these two phenomena. The aim of this thesis is to connect these two gender gaps and discover whether homophily is what causes these two phenomena.

This leads to the question of how the affiliation network of editors and articles, and the gender of both, contributes to this gender gap.

To answer this question, several subquestions need to be answered first:

RQ1 How pronounced is the gender gap in users in the used dataset?

[1] http://www.independent.co.uk/life-style/gadgets-and-tech/news/wikipedia-seeks-women-to-balance-its-geeky-editors-2333605.html

- Analyse the amount of male and female users
- Analyse the amount of edits users make in their lifetime and how this differs per gender
- Analyse how this editcount changes when only looking at the most and least active users

RQ2 How pronounced is the gender gap in person pages in the used dataset?

- Analyse the amount of person pages per gender
- Analyse the amount of edits that are made on pages and how this differs per gender
- Analyse what effect the birthyear of a person has on the gender gap in person pages

RQ3 Is homophily present in the affiliation network of editors and articles?

- Analyse the potential preferences of both genders with regards to editing pages related to a certain gender...
  - when looking at the number of edits they make
  - when looking at the time spent editing this page
  - when only looking at people that have edited a page more than a certain number of times
  - when only looking at the most edited pages from Wikipedia
- Analyse the potential preferences of both genders with regards to editing one page for a longer time (in depth) or edit several pages once (in width)

Overview of thesis This thesis will start by giving more context about Wikipedia and the different gender gaps that can be observed. After that, the used data will be outlined. Furthermore, the analysis of this data will be described in detail. The different results will be used to attempt to answer the previously named subquestions. This thesis will then end by giving the final results.

2 Related Work

Since the beginning of Wikipedia in 2001, a lot of research has been done on Wikipedia, but the last few years, research is more often focused on the gender gap that is (and always has been) present on Wikipedia. It is important to note that most research - including this thesis - focuses on the English Wikipedia. Previous research has concluded similar gender gaps can be seen across at least six different languages [Wagner et al., 2015]. Most research focuses on the English Wikipedia because it is the largest version: 93% of all readers read this version and 49% primarily read it. It also the most edited version, with 76% of all contributors editing this version and 40% primarily editing it [WikimediaFoundation, 2011b].

Studies on gender bias in Wikipedia can be divided into three groups: studies about the readers of Wikipedia, studies that look at articles of “objects” with a gender and studies concerning the editor community.

2.1 The readers

The Wikipedia community mostly consist of readers: the 2011 Readership Survey showed that the number of non-contributing readers on Wikipedia was 94%. These readers are often called free-riders, who only use information and give nothing back, and are often depicted with a negative connotation [Antin and Cheshire, 2010]. However, research has shown reading without modifying a text is also an indication of the value and reliability of an article.

Since the early days of the internet, females have been less frequent internet users than males [Bimber, 2000]. Nowadays, females are using internet almost as much as males are, but they are using it differently in comparison to their male counterparts. Females show a strong preference for usage of the internet to as a communication medium, while males are more likely to create content or read information online [Poindexter et al., 2010]. This can also be observed on Wikipedia, where 56% of the readers are male [WikimediaFoundation, 2011b]. It can thus be concluded that the gender gap in Wikipedia readers is almost non-existent, or at least very small.

Apart from the amount of males and females, another difference in males and females was found. While male students used Wikipedia more frequently and had a more positive attitude towards it, female students displayed more cautious attitudes, emotions, and behaviour: male students think higher of i.a. the accuracy of the information, the correction of inaccurate information and the writers knowledge [Lim and Kwon, 2010, Lim and Kwon, 2009] .

2.2 Person Pages

Wikipedia does not have any hard-and-fast rules, but they do have a policy and guidelines regarding its principles and best-agreed practices [Wikipedia, 2016a].

All person pages on Wikipedia must meet the notability guidelines, which state that: “People are presumed notable if they have received significant coverage in multiple published secondary sources that are reliable, intellectually independent of each other, and independent of the subject.” [Wikipedia, 2016b]

Many research papers on Wikipedia have already concluded there is a gender gap in person pages on Wikipedia: of all person pages on Wikipedia, only 15% are about women [Graells-Garrido et al., 2015].

This percentage increases when only looking at people who are born later, which could be a clue that the gender gap is slowly becoming less present [Klein, 2015]. But while part of this gender gap can be explained due to historical reasons and social contexts, these are not the only causes.

The lack of source material can be a possible cause of the gender gap in person pages. Wikipedia is based on other research materials, so if these materials have a bias, so does Wikipedia [Reagle and Rhue, 2011]. However, the bias in this original research is not large enough to cause the gender gap on Wikipedia.

This gender gap does not only present in person pages, but in all pages. Because of the way Wikipedia works, the content reflects the interest of the users. Even Wikipedia co-founder Jimmy Wales has said that Wikipedia has many articles on Linux that nobody proposes to delete, while many topics about fashion get deletion notices from editors who think fashion is unimportant [Armstrong et al., 2013]. This is an example of topic-specific entry requirements, where a ‘feminine topic’ (a topic more women than men are interested in) is more likely to be removed.

The difference between male and female person pages is not only present in the amount of pages each gender has, but also the language used etc. Research about this can be divided into different categories:

Notability Women on Wikipedia are on average more notable than men. This is presumably due to gender-biased determination of value, mainly that pages about non-famous women are more likely to be deleted than non-famous men [Wagner et al., 2016]. This leads to an even larger gender gap when looking at articles of people who are not very famous. An example of this is that a page of a acclaimed female writer [2] might only have a Wikipedia page of three paragraphs, while a character of a video game[3] has fifteen [Cohen, 2011].

This gender gap in notability is also shown by the fact that women are 14% less likely than men to have a page on Wikipedia if they are only relevant for one language edition [Wagner et al., 2016]. This is important since this group of people who are only relevant in one language edition makes up 45% of men and 40% of women.

Lexical analysis Men are seen as the null gender, which means it is often explicitly stated when one is female while this does not happen at pages about men [Wagner et al., 2015]. An example of this is that the first paragraph of a female page might contain words like ‘woman’, ‘female’ or ‘lady’. This shows that there is a clear difference in the content of pages about males and females.

When looking at the most indicative words for each gender, it can be noted that pages about women talk more often about family and relationships than pages about men do [Wagner et al., 2015]. When looking at the words that were most indicative of a specific gender, 23-32% of these are family oriented at female pages, against only 0-4% in male pages. The “spouse” attribute, which is shown in the infobox, is also more frequently used for women [Wagner et al., 2016].

Network structure Differences between men and women are also present in the network of Wikipedia persons. This network will be described in Subsection 2.4.1

2.3 Editors

When looking at all users active on Wikipedia, only 9-16% of these users is female [WikimediaFoundation, 2011a, Lam et al., 2011]. Many possible causes for this gender gap have been suggested.

When looking at Wikipedia editors, a gender gap was to be expected, since it was already stated in Section 2.1 that a small gender gap was present in Wikipedia readers. The gender gap in editors is so large however, that the

[2] The page has been edited and elongated since: https://en.wikipedia.org/wiki/Pat_Barker

[3] https://en.wikipedia.org/wiki/Niko_Bellic

gap in readers seems an unlikely reason, since the gender gap in readers is less pronounced than the gender gap on Wikipedia, suggesting multiple causes for the gender disparity [Chiu et al., 2013].

Another possible explanation can be that women have less confidence in their expertise or lower confidence in the value of their contribution. When looking at a survey of contributors to Wikipedia, women are more 43% more likely to claim they do not have enough knowledge or expertise and 22% more likely to claim they do not have enough information [Collier and Bear, 2012]. When only looking at people with similar number of years of education and partner, women are still more likely to claim they have less expertise. This observation can also be observed outside of Wikipedia: women are less inclined to have confidence in their own work or expertise than men do [Blanch et al., 2008].

Apart from this, women are also more likely to respond they do not feel comfortable editing other peoples work than men. They are 34% more likely to leave Wikipedia as an editor because of this than men and 23% more likely to give this as a reason for not being more active [Collier and Bear, 2012].

Internet spaces in general too often have a gender gap because the lack of female perspective and role models, and off-putting language. While Wikipedia has numerous good-faith norms and is aware of the need to welcome and support women, they are still off-putting to many women [Reagle, 2013]. Recent discussions on systemic gender bias on Wikipedia made it clear a number of women were not comfortable contributing to this conversation on Wikipedia, since it was not a friendly environment [Reagle, 2010]. It has even occurred that women who often fill complaints against sexist users, are banned themselves, because they reason that if you file a lot of complains, you yourself must be the problem [Paling, 2015]. A possible reason for this named in this article is because the people who decide who is banned are mostly white, formally educated males from the global north who tend to side with men.

Wikipedia has made several attempts to close this gender gap. An example is that the Wikipedia edit page has been made more visual and easier to use for people who are new to Wikipedia. Coverage of Wikipedia's gender gap has shown many commentators denied this gap was a problem. They also said it was a problem for women: it was their fault for not joining, their choice not to contribute and mocked *girly* interests [Eckert and Steiner, 2013].

The Wikimedia Foundation, an initiative started and sponsored by Wikipedia, has workshops in an attempt to attract more females. These workshops explain the gender gap and show females how to edit Wikipedia. Research has shown new female users often started writing because they heard about the gender gap and more often write about feminism or gender, while men started writing because they felt empowered to do so or had knowledge to add and more often write about their personal interests [Armstrong et al., 2013].

2.3.1 Edit history and discussion pages

Several authors have suggested the gender gap in editors could lead to a high level of conflict in the editing and writing process within Wikipedia. The survey of contributors to Wikipedia showed females were 26% more likely to say they left Wikipedia because they got into conflicts with other Wikipedians and were 31% more likely to say they were not very active because of fear of be-

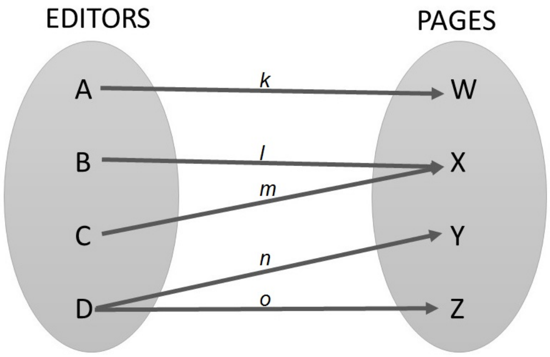


Figure 1: A graphic representation of the relations in the affiliation network

ing criticised [Collier and Bear, 2012]. This could also be a possible reason to why females tend do work in topics in which the discussion pages have a more positive tone [Laniado et al., 2012].

While the previous study suggest the behaviour of women plays a big part in the gender gap, another study actually states the gender gap is because edits by women are more likely to be reverted and not because women react differently once their edits are reverted [Lam et al., 2011]. Both studies do agree on the fact that women prefer to be active in the Wikipedia community and be social. This same result is also found in other research, that suggests women send longer messages (83 vs 71 with users and 85 vs 68 with administrators) and more often include links to the Wikipedia policies (6.2% vs 2.5% with users and 12.4% vs 4.9% with administrators) [Laniado et al., 2012].

It is interesting to note that while research is often very positive about the discussion pages, many individuals do not agree with this and 22% of women have personally had unpleasant experiences [WikimediaFoundation, 2011a].

2.4 Affiliation network

An affiliation network is a network that represents the affiliation of people (on the left) with their foci (on the right) [Easley and Kleinberg, 2010]. In Figure 1, a very simplified version of the affiliation network between users and pages can be observed. Within this network, several relations can be observed, which will be explained in the following sections.

2.4.1 Hyperlinks between pages

On Wikipedia, one person page may contain a hyperlink to another person page, this is called a inter-article link. This often means there is a direct relation between these two people. There are no strict rules for adding hyperlinks to pages, which results in some missing links, yet surprisingly few noisy links [Adafre and de Rijke, 2005]. This makes Wikipedia a platform with a rich link structure, ideal for constructing a network.

When constructing a network from these hyperlinks, it can be concluded that the top-ranked women are slightly less central than men, but this centrality of women decreases faster than that of men [Wagner et al., 2016]. Pages about women also link more to pages about men, than pages about men link to pages about women, which might be a reason why women are not as central in this network as men [Wagner et al., 2015]. This hyperlink network changes per language and while large parts of the network are similar, the most central persons often differ, but the gender gap remains [Aragón et al., 2012].

2.4.2 User has edited page

Besides the hyperlinks, which exists between two pages, there is also a relation between pages and their editors. These are the arrows that can be noticed in Figure 1. Each of these arrows also has a weight, shown by the number on the arrow. For example, editor A has edited page W k times. These arrows make up the affiliation network.

As mentioned in the introduction (Section 1), the founder of Wikipedia himself expects there to be homophily in this relationship between users and pages. This means males tend to edit pages of male persons and female tend to edit pages of female persons. If this were true, this would mean there is a relationship between the gender gap in person pages and the gender gap in users. So because women edit more about women, the shortage of women editors leads to a shortage of pages about women.

No studies were found that consider the affiliation network of editors and articles and the gender of both, which is why this thesis will focus on this network.

2.4.3 Users have edited same page

Another interesting relationship is that between users who have edited the same page. An example of this is user B and C in Figure 1, since they have both edited page X.

When two nodes in a network have a neighbour node in common, the formation of a new link between these two nodes is often called triadic closure [Easley and Kleinberg, 2010]. In our network in Figure 1, this could be a formation of a link between B and C, since they both edit page X. Triadic closure is very natural since they are more likely to come in to contact with each other, since they both edit the same page.

It could be that users that hold this relationship are similar, since they edit similar pages thus might have similar interests. Females and males focus on different broad content areas, which suggests some areas on Wikipedia will show different gender gaps [Lam et al., 2011]. This supports the claim that users that edit the same page might be similar and homophily is present in this relationship.

When looking at discussion pages of subjects, not only gender homophily, but also but emotional homophily has been found [Laniado et al., 2012]. Female editors have a preference to communicate with other female editors: the number of messages exchanged among women is much higher compared to males. Users also send and receive messages to users that have a similar emotional style or have similar average length of their comments.

Not a lot of research has been done on this relationship regarding gender, which makes it an interesting topic for further research.

2.4.4 Pages edited by same user

When pages are edited by the same user, it is expected that both these pages are of interest of this user, which could mean they are related or similar pages. When more users are editing the same two pages, the chance of them being of similar interest increases. A link forming between these users is another example of triadic closure.

This relationship is also notable for further research, since further research on this subject has not been performed as far as we know.

2.5 Homophily

Homophily is known as the principle that we tend to be similar to our friends or other people close to us [Easley and Kleinberg, 2010]. This relates not only to gender, but also to social status, race etc.

Homophily is present in many networks: a study on social networks show that men and women interact differently through social networks and that they have differences in their friends [Volkovich et al., 2014]. Here, homophily is more present in women than in men, since they accept more friend requests from other females than men.

When looking at social networks, research also shows homophily is mostly present when a person does not have many connections [Thelwall, 2009]. This homophily than becomes less noticeable as the person gains more connections. This could also be related to the fact that a smaller gender gap is seen when looking at more famous Wikipedia pages.

On social networks, the affiliation network is a bipartite graph with two types of nodes: users and groups the user is a part of. On Wikipedia, these node types are users and edited pages.

Research on social networks has shown a power law in the group size distribution, with many small groups and fewer large groups [Zheleva et al., 2009]. The same distribution is found in the activity of editors on Wikipedia and it is expected to see the same with the person pages: some person pages are edited very often and many are not [Antin et al., 2011].

3 Data

Several datasets were necessary to get all information, namely about the revision history of Wikipedia, gender and editcount of users, and gender and birthyear of people that have pages on Wikipedia.

Since research on this subject has been done before by several students of the UvA (Houda Alberts, Paul Schrijver), I did not have to do any data-scraping myself. Details about this data can be found in Section A.1. As can be seen in Figure 2, five datasets were provided, which were then merged into two final datasets. The merging of these datasets meant a lot of data had to be removed, because no overlapping features could be found.

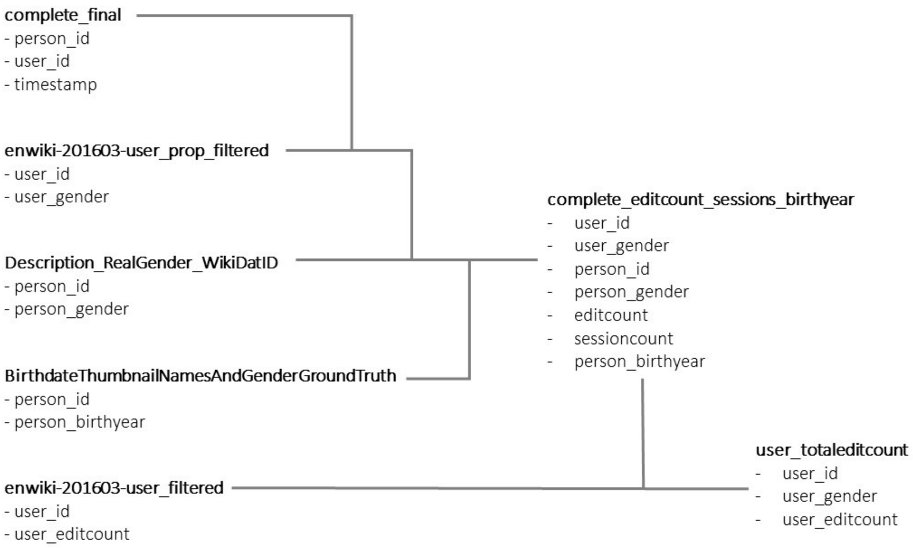


Figure 2: A graphic representation of the merged datasets and their results

	Editors	Pages	Outgoing arrows	Incoming arrows
Male	59.812	171.487	2.300.756	2.068.117
Female	6.352	31.890	209.328	441.462
Total	66225	203377	2.510.084	2.51.0084

Figure 3: Overview of the content of Edits dataset

3.1 Edits

The first dataset, `complete_editcount_sessions_birthyear`, contains all edits of which the gender of both the user and person page was known, as well as the person page birthyear. This means more than 5 million edits are present. These edits are saved using the combination of the user and page and their genders.

The original dataset with all users and their gender contained 103.694 females and 525.170 males. The dataset with the person genders had 127.947 females and 731.236 males. Table 3, which contains the cardinality of the users and person pages of the merged dataset, shows that only a very small percentage of these original datasets is present in our merged dataset. Only around 10% of all users of the original dataset were also in the merged dataset and only 19% of the persons. It is also important to note that of our original dataset, only 6% of female users were in the merged dataset, while 10% of the male users were. For the dataset of editors this was more equal with 19% of male person pages and 20% for females.

To make this merged dataset smaller, edits with the same user-person combination are merged and the amount of edits this user has made on this page is saved in `editcount`. This shrinks the dataset to about 2,5 million rows. Be-

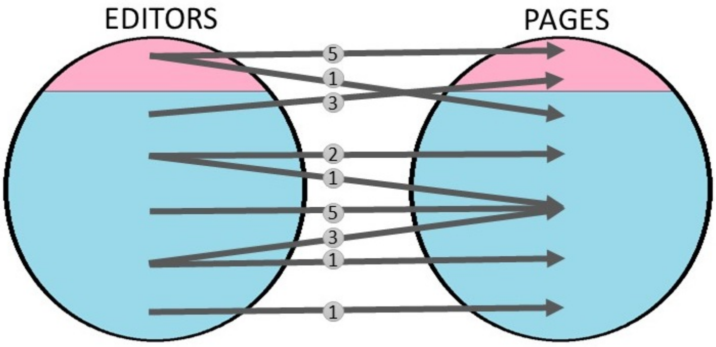


Figure 4: A graphic representation of the affiliation network of Wikipedia, with male and female editors, male and female pages, and the weighted relations between them

cause some users make a lot of edits in a row, `sessioncount` was added. The sessioncount shows how many sessions of editing a person has done on this page. One session is counted as a period of time where there was less than 24 hours between each edit. If a person thus edits one page 30 times on one day, the editcount is 30, but the sessioncount is only 1.

This dataset makes up the affiliation network with the user on one side and the persons on the other side. Each row in this dataset represents one arrow in our network, with the editcount and sessioncount being the weight of that arrow.

Figure 4 shows a graphic representation of this network. The two circles represent the group of editors and the group of pages, both of which have male and female persons. Arrows with a certain weight, either the editcount or sessioncount, go from editors to pages and show who has edited who and what their genders are. One page might have several incoming arrows, this means many people have edited this page. If an editor has many outgoing arrows, this means it is an active editor who has made edits on several different pages.

3.2 Users

The other dataset, `user_totaleditcount`, shows the amount of edits a certain user has done in its lifetime. It contains only the users from the previous dataset whose total editcount was known. It also provides us with the gender of this user. Figure 5 shows the distribution of the amount of edits made per gender. One can see men have made a larger number of revisions and that there are more men than women in this dataset, with a total of 6.310 females and 59.505 males.

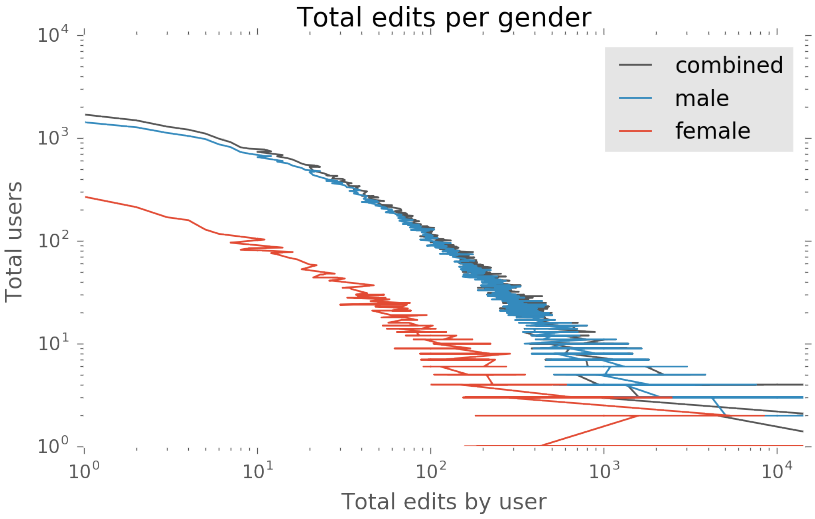


Figure 5: Distribution of amount of revisions per gender on log-log scale

4 Analysis

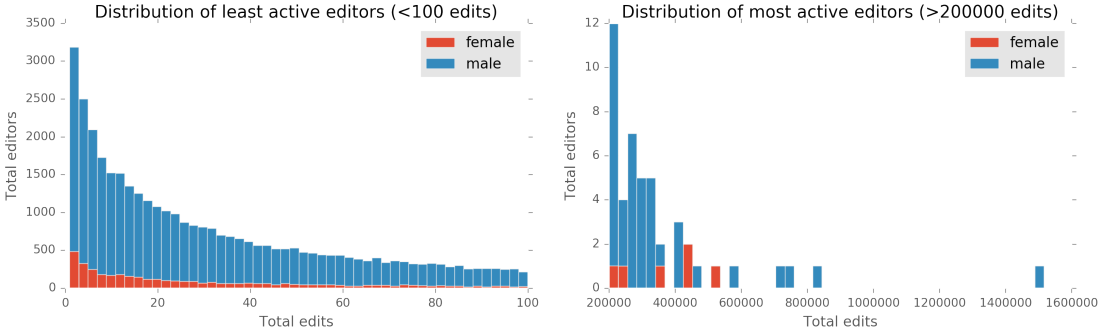
4.1 Gender gap in users

To plot information about users, `user_totaleditcount` was used to find a relation between the amount of edits a user has made in its lifetime and the gender of this user.

When looking at the amount of edits made, regardless of gender, we can see most persons only make a few edits: almost half of all users have made 100 edits or less in their lifetime. In Figure 6a, the distribution of these users can be observed. It is clear many users are only making a few edits and then stop editing at all. The gender gap seems relatively stable against the amount of edits made by a person: when only looking at people who have made less than 100 edits, the percentage of female users is 10,9%, only 1,3% higher than average. Figure 7 shows that this percentage continues to change when looking at different amounts of edits made.

Figure 8a displays the average and median editcount of both genders and the amount of editors in our dataset. Our dataset contains 9,6% female users, which is similar to the 8,5% found by Wikipedia, but a much larger gap than the 16% found in other papers [WikimediaFoundation, 2011a, Graells-Garrido et al., 2015]. Apart from this gap in the amount of editors, we also see that men on Wikipedia edit 1,2 times more than women when looking at averages and this number grows to 1,56 when looking at means. This leads to a percentage of only 8,0% female edits in this dataset.

Previous research has stated the bottom 75% of men make a similar number of edits as the bottom 75% of women [Antin et al., 2011]. As Figure 8b shows, we did not find this. In our dataset, we do have a small percentage of women and the average woman made fewer edits than the average male. This was also true when not taking the average, but the median.



(a) Distribution of the least active users (less than 100 edits: 36375 users)

(b) Distribution of the most active users (more than 200.000 edits: 47 users)

Figure 6: Distribution of the amount of edits made against the amount of users who have made this amount of edits

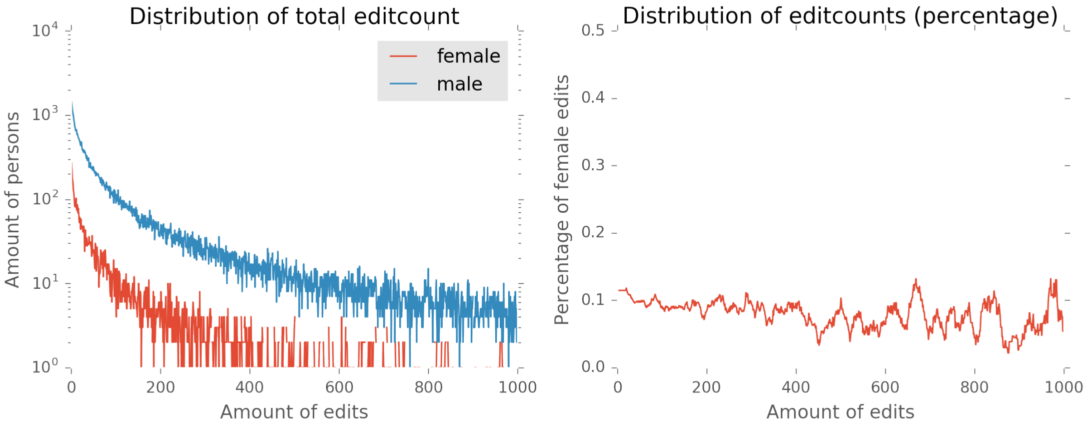
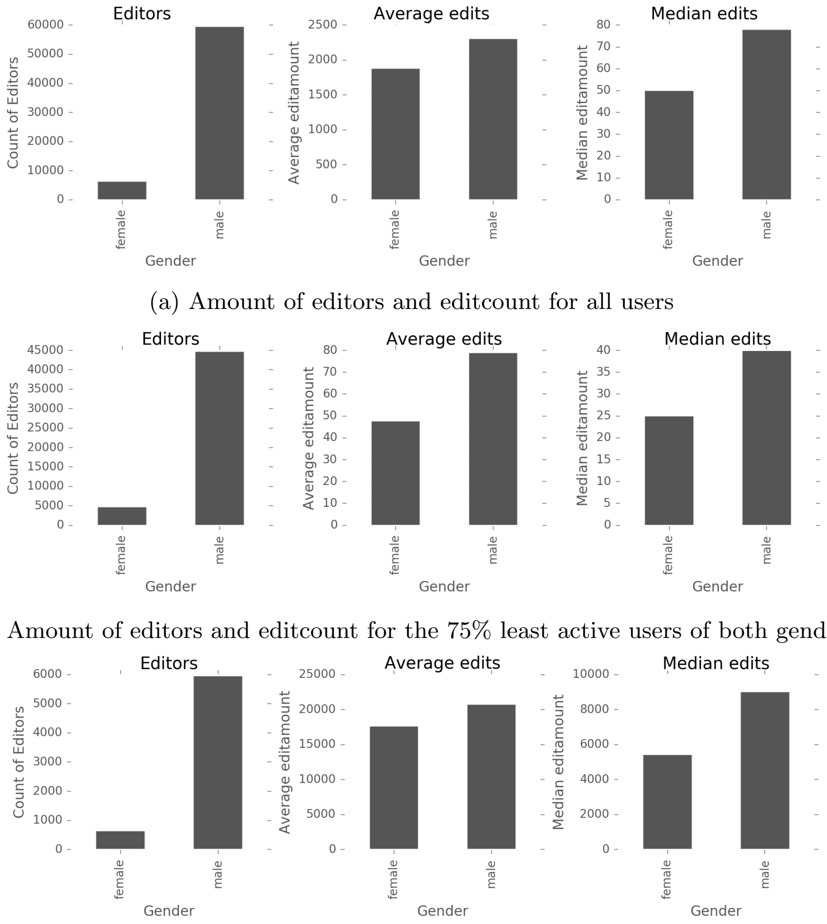


Figure 7: Gender gap plotted to the total editcount of a user, in amount of users and percentage females



(a) Amount of editors and editcount for all users

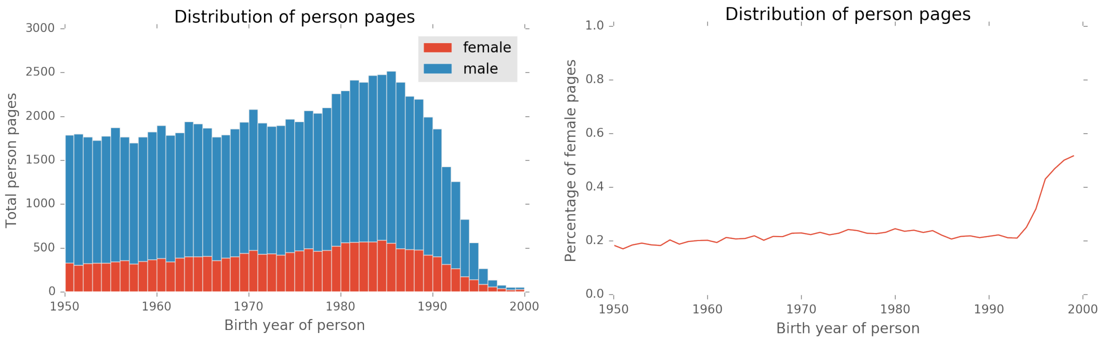
(b) Amount of editors and editcount for the 75% least active users of both genders

(c) Amount of editors and editcount for the 10% most active users of both genders

Figure 8: Average and median amount of edits made for each gender when looking at different subsets of data

The same was found in the 75% least active editors, but this gave an even larger gap in the amount of edits made when looking at the average. The median was around the same when looking at the full dataset, which means a gender gap was still present.

It is worth noting that the gender gap in average edits of Figure 8c was actually the smallest gender gap found. Even when looking at the median amount of edits instead of the average, it is clear that this gender gap is smaller than the one found in the least active users instead of the way around. The most active person on Wikipedia (who is responsible for almost 1.5 million edits) is male, but when looking at the top 10 of most active users, 3 of these are female.



(a) Amount of males of females on Wikipedia, sorted by birthyear

(b) Percentage of female pages on Wikipedia, sorted by birthyear

Figure 9: Gender gap in person pages from 1950

4.2 Gender gap in person pages

When looking at the dataset `complete_editcount_sessions_birthyear`, where the person pages and their gender are stored, it is found that 15,7% of all person pages are about females. Of all edits made, 17,6% is made on female pages, which means the edits-per-page ratio of female pages if slightly higher than that of men.

Previous research stated there was a relation between the gender gap and birthyear, and that the gender gap in pages about younger people would be smaller if not nonexistent [Graells-Garrido et al., 2015, Klein, 2015]. Figure 9b show the percentage of female pages per birthyear and the same as in the literature is found: in 1993 a large spike in female pages is found. This corresponds with the findings from Graells-Garrido et al. (2015), but this does not mean this spike will stay as time passes. It could be that the gender gap is not present because the gender gap only appears in adults or that not enough data is found.

But this does not mean the gender gap is not getting smaller as the birthyear gets higher. Figure 10 shows the percentage of pages that have people born in a certain year for both men and women. This plot shows us that for men, less than 1% of all pages are about someone with a birthyear 1980, while this percentage is around 1,5% for women. This is another marker that pages about females are often about people with a later date of birth.

These two findings mean that the gender gap on Wikipedia is related to the birthyear. Further research could determine whether this is because of better historical sources or the real life gender gap.

4.3 Gender gap in affiliation network

When looking at the affiliation network introduced in Section 3, it is important to know who is editing who. Figure 11a shows what percentage of arrows in our network is going from a certain gender to another gender. So this shows that of all users that have edited pages, 82% of these pages are about males. Something very interesting is that this number is the same for male and female

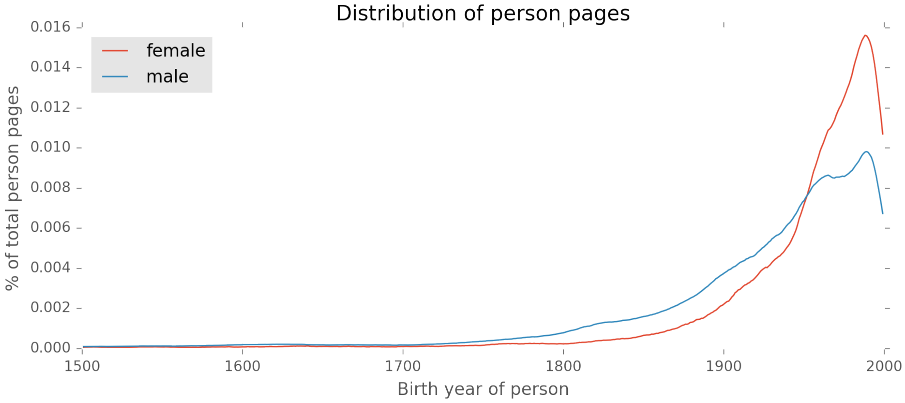
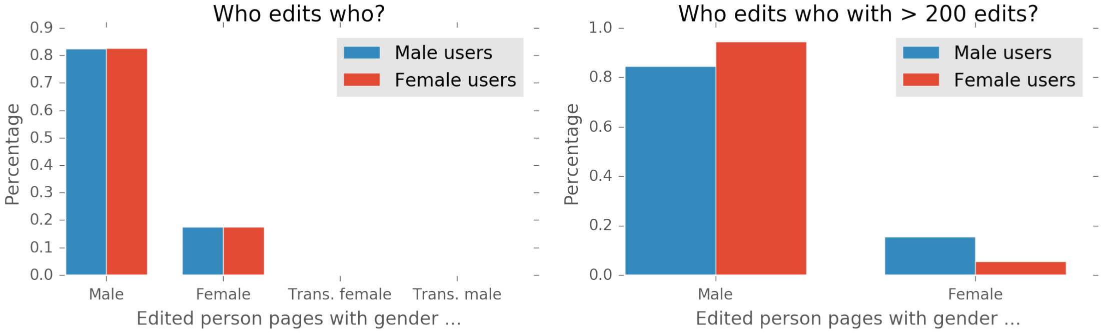


Figure 10: The percentage of pages from people born in a certain birthyear



(a) Percentages when looking at all arrows in the network

(b) Percentages when only looking at arrows with a weight of more than 200

Figure 11: The percentage of female and male pages edits for each gender

users. This means users do not have the preference of editing someone of their own gender, so homophily is not present when measuring this way.

When only looking at arrows of a certain weight, so for example only at users and pages where a user has edited a page more than 200 times, the preference of a user changes slightly, but not in the way expected. When looking at arrows with a weight of more than 200, females actually show a preference of editing pages about males, which means inverted homophily is found.

Figure 12 shows when this inverted homophily is present by plotting the percentage of male pages edited against the minimum width of the arrows (the editfilter). This shows us males are very consistent in their editing: the amount of male edits always stays around 82%. But when looking at female users, their editing behaviour changes as the editfilter changes.

The distribution of editors based on editfilter is a bit different for men and women. Figure 13 contains two plots, both of which show the percentage of a certain editcount of sessioncount. We can see that of the total number of edits by females, the percentage of large editcounts is a bit lower than with males.

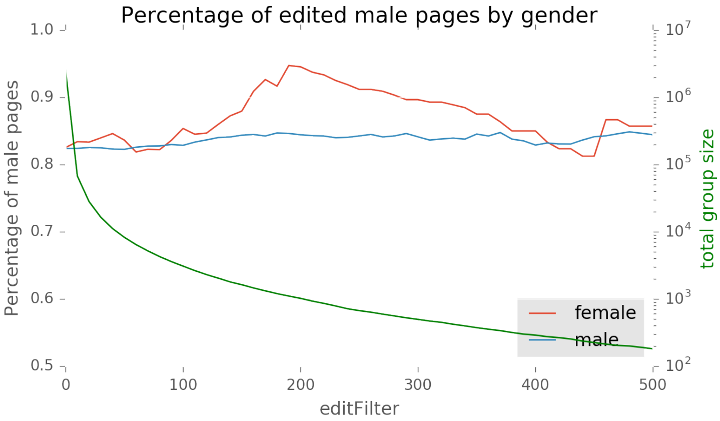


Figure 12: The difference in edit preference shown per minimum weight, shown together with the groupsize of the data

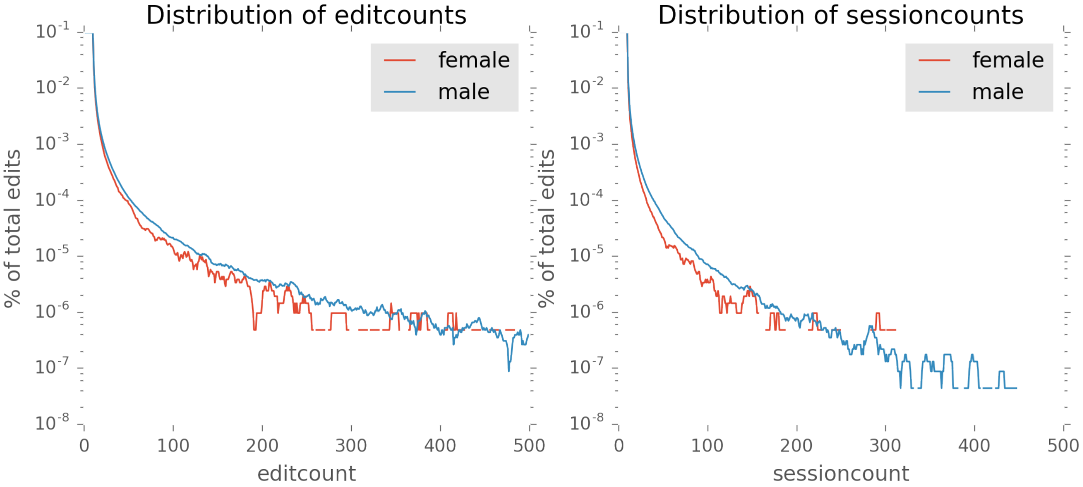


Figure 13: Male and female editors and their sessions and edits on a logarithmic scale

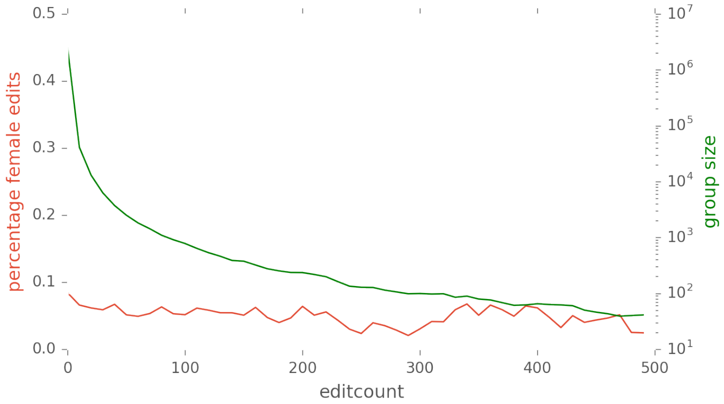


Figure 14: Difference in men and women in how much they edit, together with the groupsize of the data. Percentage of female editors was taken over increasingly large groups of editors to keep the groupsize as high as possible

The same can be said for sessioncounts. Both genders mainly edit pages a low number of times with a low number of sessions, but sometimes pages are edited more often. Sometimes gaps appear in this plot, especially when looking at female sessioncounts. This is because this data is sparse.

The same decline in female pages can be seen when looking at the percentage of females for each editfilter in Figure 14: the percentage of females decreases as the editfilter increases. At higher editfilters, the percentage of women sometimes is quite high, but this is because not much data is available, so one female editor might already cause the percentage to spike 50% if only 2 editors with this editfilter are in our dataset.

Finally, Figure 15 shows the sessioncount plotted against the editcount, so one can see the relation between these. It can be observed from the first plot that arrows with a high editcount and sessioncount are mainly from male users, but some are pages about females as can be seen in the second plot.

5 Conclusions

To find out how the affiliation network of editors and articles and the gender of both contributes to this gender gap, several subquestions were asked:

RQ1 How pronounced is the gender gap in users in the used dataset?

RQ2 How pronounced is the gender gap in person pages in the used dataset?

RQ3 Is homophily present in the affiliation network of editors and articles?

In this dataset, 9,6% female users were found. This shows a clear gender gap in our dataset. These users also made less edits on average, which means only 8% of all edits were made by females. Even when only looking at the least active 75% users of each gender, a gender gap in the average and median amount of edits was found. This gender gap was least present in the 10% most active users, something that contradicts previous research [Antin et al., 2011].

When looking at person pages, a gender gap of 15% was found. This gender gap decreases when looking at people with a later birthyear. The gender gap

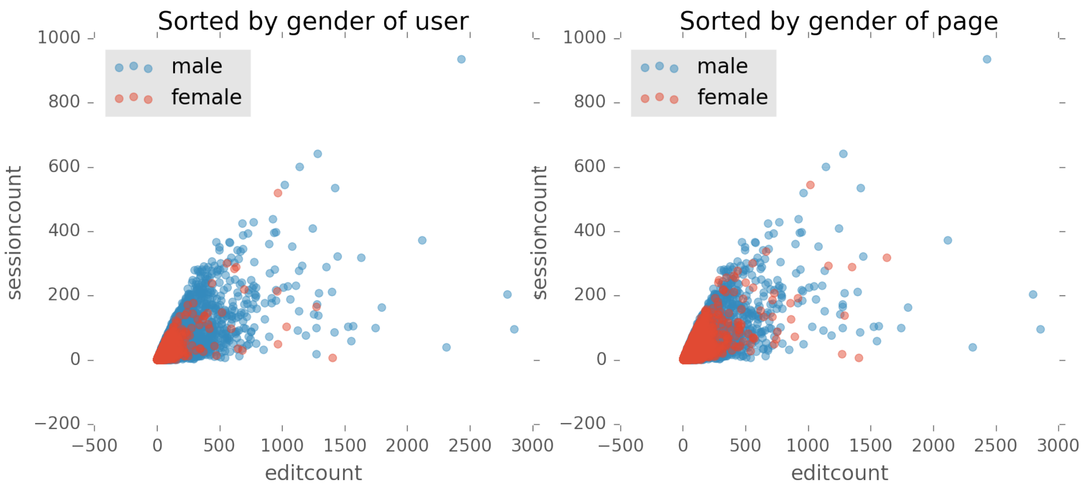


Figure 15: Difference in men and women in the editcount-sessioncount ratio

declined to more than 30% when looking at people born after 1993, but not enough data is available to draw conclusions from this.

Furthermore, the affiliation network showed differences in male and female users are not as pronounced as expected. Homophily is not found in edit preference, while this was suggested by many. Inverted homophily was even found when looking at female edited with more than 200 edits on a single page. Males had a slight preference for making more edits on a single page than females, but this difference was not very large.

To conclude, a gender gap in Wikipedia was found in the person pages and in the users, but these gaps can not be explained by homophily between users and persons.

5.1 Future work

While much research has already been done on Wikipedia, there are still gaps in our knowledge. This is because most research focuses on the gender gap in the person pages and the gender gap in users, but does not combine these two to come to further insights.

A relationship not often explored is that between two users that have edited the same page. As said in Section 2, it is probable that these two users have things in common and it would be interested what that is. The other unexplored relationship is that between two pages that have been edited by the same user, which is interested because of the same reasons. This relationship between three people is often called a triangle.

Something worth reviewing is single gender triangles. A single gender triangle is a connection between three people of the same gender. These connections may be one-sided, but there has to be a connection between each vertex. These single gender triangles are especially common in social media networks, where they are much more abundant than mixed gender triangles, especially for male users [Volkovich et al., 2014]. It would be interesting to see if these also occur more frequently when looking at the Wikipedia network.

While several papers talk about the gender gap in person pages declining after 1990, not many give an explanation for this. It would be good if research could show whether this is because of the age of these people, the declining gender gap etc. If this was researched, this would also give new knowledge to why the gender gap in person pages might exist and how to make it smaller.

Something new discovered in this thesis is the fact that male and female users are very similar in which gender they edit, so no homophily has been found. But this changes when only looking at people who have edited a single page a lot of times, then slight inverted homophily is present. Finding out whether this also occurs when using a larger data set, would be another good subject for further research.

It could also be that homophily (or lack of homophily) is present in other forms of services that provide information, like other encyclopedias or news agencies.

This research also does not look at the size of an edit, which means all edits are equal, whether they are a spelling correction or the adding of several new paragraphs. This is why the parts of this research regarding the affiliation network could be further investigated using another, more detailed dataset.

5.2 Discussion

While my data contained a few million edits, this is only a fraction of all edits, editors and pages on Wikipedia. Many merges on the original datasets were necessary to get the affiliation network, which meant a lot of data was lost. It would be better if we could have made the original datasets more overlapping, so more usable data was present.

Also, as noted in Section 3, the merge I took did not have the same ratio in male and female users as my original dataset. This is another reason why it is important to redo this research with more data.

Because this research only includes editors who have registered their gender, there might be a common trait between them that influences their editing. Females might hide their gender more often than men, because they do not want to attract unwanted attention. There is also no way to check if the registered users registered their true gender and if they are the only ones working on their account [Antin et al., 2011].

5.3 Acknowledgements

First, I would like to thank my supervisor, Maarten Marx, for helping me every step of the way. I would also like to thank Paul Schrijver, who wrote his thesis “Gender gap on Wikipedia: visible in all categories?” and was willing to share his data and to help me with any problems that occurred in the merging process.

References

[Adafre and de Rijke, 2005] Adafre, S. F. and de Rijke, M. (2005). Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, pages 90–97. ACM.

[Antin and Cheshire, 2010] Antin, J. and Cheshire, C. (2010). Readers are not free-riders: reading as a form of participation on wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 127–130. ACM.

[Antin et al., 2011] Antin, J., Yee, R., Cheshire, C., and Nov, O. (2011). Gender differences in wikipedia editing. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 11–14. ACM.

[Aragón et al., 2012] Aragón, P., Laniado, D., Kaltenbrunner, A., and Volkovich, Y. (2012). Biographical social networks on wikipedia: a cross-cultural study of links that made history. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 19. ACM.

[Armstrong et al., 2013] Armstrong, C. L., Andsager, J. L., Antunovic, D., Bissell, K., Brown, T., Butler, S., Byerly, C. M., Bystrom, D., Collins, S. J., Davis, D. Z., et al. (2013). *Media disparity: A gender battleground.* Lexington Books.

[Bimber, 2000] Bimber, B. (2000). Measuring the gender gap on the internet. *Social science quarterly*, pages 868–876.

[Blanch et al., 2008] Blanch, D. C., Hall, J. A., Roter, D. L., and Frankel, R. M. (2008). Medical student gender and issues of confidence. *Patient education and counseling*, 72(3):374–381.

[Chiu et al., 2013] Chiu, S.-I., Hong, F.-Y., and Chiu, S.-L. (2013). An analysis on the correlation and gender difference between college students internet addiction and mobile phone addiction in taiwan. *ISRN Addiction*, 2013.

[Cohen, 2011] Cohen, N. (2011). Define gender gap? look up wikipedias contributor list. *New York Times*, 30(362):1050–56.

[Collier and Bear, 2012] Collier, B. and Bear, J. (2012). Conflict, confidence, or criticism: An empirical examination of the gender gap in wikipedia. In *CSCW12: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 383–392.

[Easley and Kleinberg, 2010] Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.* Cambridge University Press.

[Eckert and Steiner, 2013] Eckert, S. and Steiner, L. (2013). (re) triggering backlash: Responses to news about wikipedias gender gap. *Journal of Communication Inquiry*, 37(4):284–303.

[Graells-Garrido et al., 2015] Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 165–174. ACM.

[Hill and Shaw, 2013] Hill, B. M. and Shaw, A. (2013). The wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one*, 8(6):e65782.

[Klein, 2015] Klein, M. (2015). Wikipedia in the world of global gender inequality indices: what the biography gender gap is measuring. In *Proceedings of the 11th International Symposium on Open Collaboration*, page 16. ACM.

[Lam et al., 2011] Lam, S. T. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., and Riedl, J. (2011). Wp: clubhouse?: an exploration of wikipedia's gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 1–10. ACM.

[Laniado et al., 2012] Laniado, D., Kaltenbrunner, A., Castillo, C., and Morell, M. F. (2012). Emotions and dialogue in a peer-production community: the case of wikipedia. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 9. ACM.

[Lim and Kwon, 2009] Lim, S. and Kwon, N. (2009). Gender perspective, information behaviors, and wikipedia. *Proceedings of the American Society for Information Science and Technology*, 46(1):1–5.

[Lim and Kwon, 2010] Lim, S. and Kwon, N. (2010). Gender differences in information behavior concerning wikipedia, an unorthodox information source? *Library & information science research*, 32(3):212–220.

[Paling, 2015] Paling, E. (2015). Wikipedia's hostility to women. `http://www.theatlantic.com/technology/archive/2015/10/how-wikipedia-is-hostile-to-women/411619/`. [Online; accessed 9-June-2016].

[Poindexter et al., 2010] Poindexter, P., Meraz, S., and Weiss, A. S. (2010). *Women, men and news: Divided and disconnected in the news media landscape*, chapter Women, Technology, and News. Routledge.

[Reagle and Rhue, 2011] Reagle, J. and Rhue, L. (2011). Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21.

[Reagle, 2010] Reagle, J. M. (2010). *Good faith collaboration: The culture of Wikipedia.* MIT Press.

[Reagle, 2013] Reagle, J. M. (2013). Free as in sexist? free culture and the gender gap. *First Monday*, 18(1).

[Thelwall, 2009] Thelwall, M. (2009). Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231.

[Volkovich et al., 2014] Volkovich, Y., Laniado, D., Kappler, K. E., and Kaltenbrunner, A. (2014). Gender patterns in a large online social network. In *Social Informatics*, pages 139–150. Springer.

[Wagner et al., 2015] Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. (2015). It's a man's wikipedia? assessing gender inequality in an online encyclopedia. *arXiv preprint arXiv:1501.06307*.

[Wagner et al., 2016] Wagner, C., Graells-Garrido, E., and Garcia, D. (2016). Women through the glass-ceiling: Gender asymmetries in wikipedia. *arXiv preprint arXiv:1601.04890*.

[WikimediaFoundation, 2011a] WikimediaFoundation (2011a). Wikipedia editors study. `https://upload.wikimedia.org/wikipedia/commons/7/76/Editor_Survey_Report_-_April_2011.pdf`. [Online; accessed 13-April-2016].

[WikimediaFoundation, 2011b] WikimediaFoundation (2011b). Wikipedia readership survey. `https://meta.wikimedia.org/wiki/Research:Wikipedia_Readership_Survey_2011/Results#Wikipedia_has_slightly_more_male_readers_than_female`. [Online; accessed 12-June-2016].

[Wikipedia, 2016a] Wikipedia (2016a). Wikipedia: Biographies of living persons. `https://en.wikipedia.org/wiki/Wikipedia:Biographies_of_living_persons`. [Online; accessed 18-June-2016].

[Wikipedia, 2016b] Wikipedia (2016b). Wikipedia: Notability (people). `https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)`. [Online; accessed 18-June-2016].

[Zheleva et al., 2009] Zheleva, E., Sharara, H., and Getoor, L. (2009). Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016. ACM.

A Datasets

A.1 Received datasets

enwiki-201603-user_filtered size: 3 x 27.618.706
This data is freely available from the Wikimedia API

Name	Type	Description
user_id	int	id-number of user
user_registration	int	registration date
user_editcount	int	how many edits this user has made

enwiki-201603-user_prop_filtered size: 3 x 628.865
This data is freely available from the Wikimedia API

Name	Type	Description
up_user	int	user_id
up_property	String	this always contained the word “gender”
up_value	String	the gender of the user, mainly male or female

complete-final size: 5 x 91.722.219
This data is freely available from the Wikimedia Foundation

Name	Type	Description
pid	int	person_id
usid	int	user_id
revid	int	revision_id
timestamp	int	when revision was made (in seconds)
minor	boolean	whether the user stated the edit as minor

BirthdateThumbnailNamesAndGenderGroundTruth size: 7 x 860.749
This data was provided by Maarten Marx

Name	Type	Description
wikiPageID	int	page_id
WikiDataID	float	person_id
birthDate	String	full birthdate
birthYear	int	birthyear
thumbnail	String	link to picture
rdf-schema#label	String	full name
GenderGroundTruth	String	gender of person the page is written about

Description_RealGender_WikiDatID size: 3 x 969.997 *This data was provided by Maarten Marx*

Name	Type	Description
WikiData_ID	int	person_id
GroundTruthGender	int	gender of person the page is written about
description	String	the category of the person

A.2 Created datasets

user_totaleditcount size: 3 x 65.815

Name	Type	Description
user_id	int	
user_gender	String	
user_editcount	int	

complete_editcount_sessions_birthyear size: 7 x 2.510.084

Name	Type	Description
user_id	int	
user_gender	String	
person_id	int	
person_gender	String	
editcount	int	the amount a certain user has edited this page
sessioncount	float	the amount of sessions the user has edited this page
person_birthyear	int	