Computational Social Science project

The primary goal of the paper here analyzed is to assess the presence of an asymmetry between men and women in Wikipedia. With the aim of increasing the awareness about it and in general the various ways gender differences can manifest online. And after having demonstrated the presence of such biases, via three methods to assess the difference of notability required to get into Wikipedia, the presence topical and linguistic bias, and structural properties differences, the relevance of the attributes on women's and men's Wikipedia pages and the centrality of women in hyperlink network formed by people's Wikipedia articles, and then to try to propose some ways to mitigate these gender asymmetries in order to account the lower visibility and the biased use of language and topics for women on Wikipedia and on the web.

To understand if there is a bias in the global notability between women and men, a study was done to see if it is harder for a woman to have her own page on Wikipedia than for a man, the method used was to look at difference in gender gap between local heroes, so people with low levels of notability, (included in just a few number of language editions of Wikipedia) and superstars, (people included in multiple language editions of Wikipedia), to conduct this study a negative binomial regression has been fitted with the number of language editions in which the person is present as dependent variable and the gender as independent variable.
To make sure that this was not just a Wikipedia problem but a problem of the web in general, a similar study has been conducted on the Google trend data as an external proxy of the popularity of a person.

Another study conducted to assess the asymmetry between women and men was an investigation to understand it there is a topical and linguistic bias in the way women are depicted, for example topical bias is done by emphasizing more on the fact that they are women or mentioning their family, while linguistic intergroup bias (LIB) states that there is a difference in the way we generalize success and instead treat failures as isolated cases in describing people in our in-group.
To investigate the presence of topical bias gender, relationship and family topics have been analyzed in the overview of biographies about men and women in their English page on Wikipedia, looking for words related to these topics and then comparing the results of the two different genders.
Regarding linguistic bias, they looked at the use of more abstract words such as adjectives rather than more concrete words to describe positive and negative aspects. In this case it was expected that there are more abstract terms to describe positive aspects in men than in women, and conversely that there are more abstract terms to describe negative aspects in the biographies of women than in those of men.

Lastly it has been investigated structural properties, since they increase the gender bias in the visibility and reachability of articles of notable women and men, to do that first they identified meta-data attributes in the dataset to understand whether there is an asymmetry in attribute distributions according to gender, by counting them and then using chi-square tests with male proportion of the attribute presence as baseline to compare it with the female proportion.
Subsequently they built a network of biographies with the hyperlink structure among Wikipedia articles, and then on this network they investigate whether there is a bias between the gender on the connectivity between people, and if the centrality of people is influenced by their gender by computing the PageRank of articles about people. Then to better understand if the observed bias

go beyond what they would have expected to observe by chance they compared their results with the results obtained from three baseline graphs, a random graph obtained by shuffling the edges of the original network, a graph obtained by shuffling the structure of the original network, preserving in-degree and out-degree sequences and an undirected small world graph using the model of Watts and Strogatz that interpolates a random graph and a lattice preserving average path length ad clustering coefficient, and then by assigning to each node a random gender maintaining the proportions of the original network.

Regarding the Wikipedia entry point as a glass ceiling, making it harder for women to be included, it has been found that the hypothesis is confirmed since for people with low or medium level of notability the gender gap is larger, indeed women born after 1900 are 10% less likely to be included than what they expected by chance and 11% for women born before 1900.
The study conducted with the negative binomial regression models highlighted how women are 13% more notable than their male counterparts, with a 12% for women born since 1900, while for women born before 1900 are 4% less notable than men, but this can be explained by the historical exclusion of women. This experiment also showed a possible further difference based on types of professions, but to confirm these would have to be done so as not to confuse this possible difference with different birth dates.
(Furthermore, it shows that the more historic a person is usually more notable, because to be remembered people have to be more important)
The study on the Google search trends of a random sample of men and women in Wikipedia born since 1900 confirmed how the women in the sample are searched in more regions and during more months. But the auctors of the paper cannot exclude that there are other confounders, for example people born recently may be more of interest on Google and women included on Wikipedia tend to be born in recent years.

The study conducted on words and their use in Wikipedia biographies partially confirmed their initial hypothesis, both for the possible topical bias and for linguistic bias that could appear in them.
For the first it has been possible to demonstrate how the distribution of words related to categories gender, relationship and family are more used for women than for men only for people born before 1900, while it could not be possible to demonstrate the same for people born since 1900.
For the Linguistic bias, the results confirmed how, as predicted by the LIB, men's biographies are 9% more likely to contain more abstract terms for positive aspects and women's biographies are 1.62% more likely to have more abstract terms for negative aspects.

Inequalities in the structure of information in Wikipedia have influence in many aspects online.
The experiment conducted on Meta-data on the DBpedia dataset showed more how there is a difference in the activities performed by men and women, with, for example, a large number of men in the sports field and a large number of women in the art world. Probably the only attribute reported in the article of which there is a topical gender difference is the attribute "spouse" significantly used more in women's biographies.

Lastly the hyperlink network has been constructed, with 700 thousand nodes, and 4 million edges, and the three baseline graphs too with the same number of nodes. With the network it is unveiled how top 30 women according to their PageRank are less central than the top 30 men. The number of women ranked among the top biographies by centralities scores indicates how they are disadvantaged in ranking algorithms, in particular when it comes for women born since 1900.

Wrapping up the results were that there is a glass-ceiling effect making it more difficult for women to be included in Wikipedia and the reason for this must be better understand and mitigated in order to have a gender equality. In the same way topical and linguistic asymmetries founded show how editors should pay more attention to do not report the bias present in other sources, and how to mitigate this tendency it could be enough to include explicitly guidelines to address gender bias. Regarding structural inequalities, although results suggest that a good job is done in interlinking women's articles, their visibility is still lower than expected, and thus there is a bias favoring men's article. For this reason, in the paper is suggested that new search and ranking algorithms should be developed to account this potential discrimination of minority groups.

A possible further study being able to have all the data possible and all the computational power needed, could be to do, for example, the study suggested in the paper itself concerning adding the relationship between date of birth and jobs to see if the greater difficulty in getting into Wikipedia for women varies according to the type of activity they do.

A study can be done on editors to understand how much gender actually influences whether people of their own in-group/gender write Wikipedia articles, and with more precise analysis how many males and how many females wrote their own Wikipedia page.

Just as a study could be done on the bias present on other web sources, one could, for example, go and look at the top 3 most visited pages on Google regarding biographies (Wikipedia excluded) or the top 3 pages found in the references of biographies on Wikipedia to see if they already have gender differences that might go some way to explaining why similar differences exist on Wikipedia.

Another interesting study would be to do such studies on different language editions of Wikipedia, possibly from different cultures to see if this gender bias is also present in them.

Additional studies

In addition to the studies done, other studies have been added:

1. Study on the number of links in the network between different genders to compare with the general case to see if there is homophily
2. Comparison to see if there are more male local heroes or not, to see if for a woman is harder to get included on Wikipedia, also a study has been done see in how many languages males and females are included, a large difference could be an indication of how differently males and females are considered, furthermore, the total number of languages in which males and females are present has been counted, so that by then dividing by the total number of males and females in the dataset we get the average number of languages in which a person of one gender is included
3. Investigation of the topical bias between genders, with focus on the number of gendered words and terms regarding personal life and comparing them with the numbers of professional terms in their summaries
4. To see the centrality of people of different genders visually, the genders were colored differently in a network structure (first including NA gender too and then without)
5. Study to find the most central nodes in the network, visualizing them in the network, with the top 10 nodes by centrality, with the color of their gender with descending size from most central to least central and all other nodes in green, and seeing the list with genders.

These additional studies have been conducted on the politicians' dataset (2793 total samples, 386 females, 2835 males, and 3221 "NA"), on a journalists' dataset (6996 total samples, 1404 females, 3661 males, and 1931 "NA"), on a dataset of Italians born after 1970 (717 total samples, 135 females, 292 males, and 290 "NA"), and on one of Italians born after 1950 in this case looking at their Italian Wikipedia page (1188 total samples, 211 females, 574 males, and 403 "NA"). (Regarding the additional datasets, the same studies as in Task B have been also done).

In addition, for the latter two datasets, since they did not concern specific occupations, it was possible to conduct a study regarding different occupations for different genders, while it was not possible to look for specific professional words for the same reason.
A study of Italian writers was also attempted, but the available dataset was too small to give statistically satisfactory results.

Results:

Politicians:

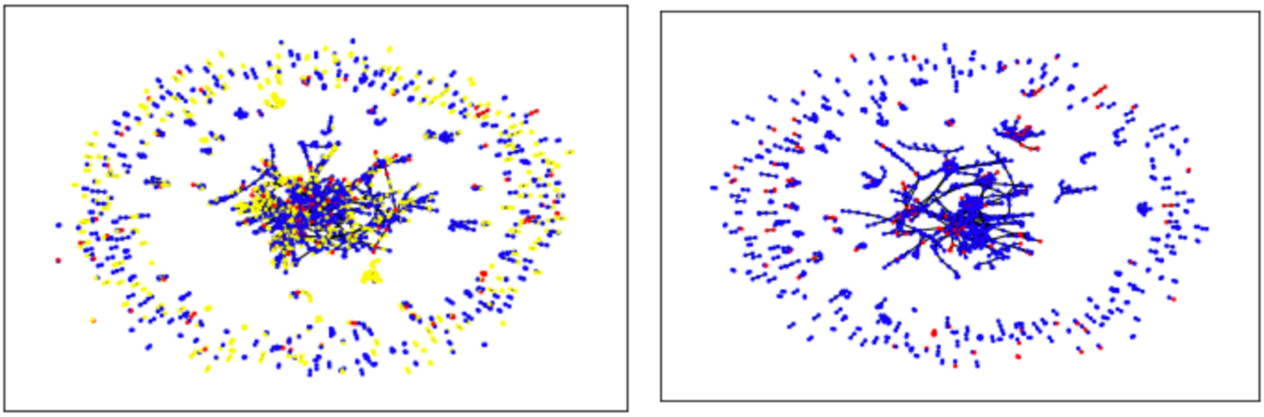
1. Here the results showed how there are just 1516 links between males and females while there are a total of 8093 links, we can say then that there are many more links between the same gender than between the different genders
2. Comparison in this case showed an almost 2% difference (17.6% for males and 15.0% for females) in the percentage of "local heroes" in people of that gender, confirming the initial hypothesis, a big difference in the total number of languages between the two genders can be seen (269 for males, 157 for females) but this can be related to the difference in the number of samples too, while for the average languages for a page there is a difference of

0.7 (almost a language more in average for women) with 6.7 languages in average for women and 6.0 for men

3.

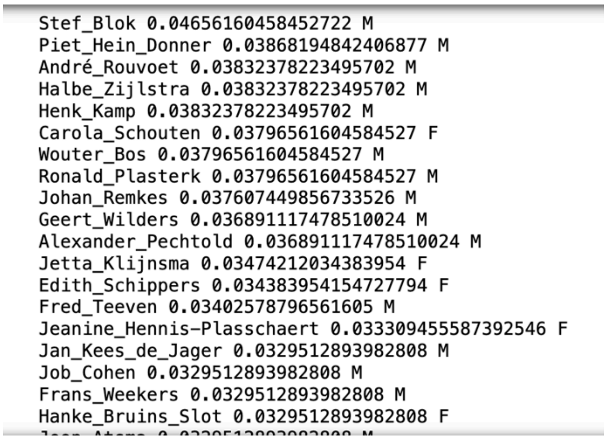
```
female gender words in politicians:  1324
female gender words average in politicians:  3.430051813471503
female personal life words:  62
female personal life words average:  0.16062176165803108
female professional life words:  458
female professional life words average:  1.1865284974093264

male gender words:  6927
male gender words average:  2.4433862433862434
male personal life words:  517
male personal life words average:  0.18236331569664904
male professional life words:  3248
male professional life words average:  1.145679012345679
```

4.

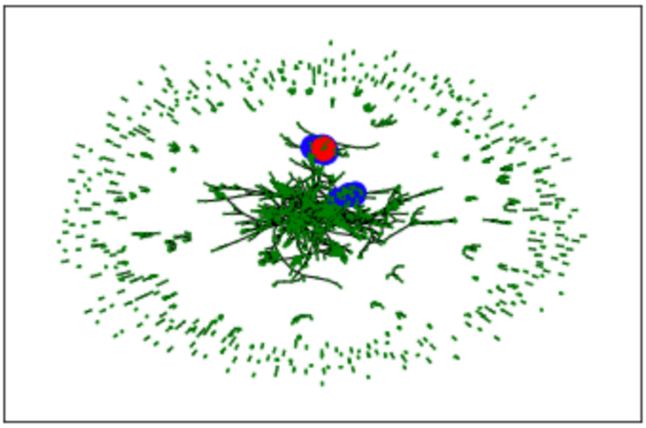


On the left the network including the “NA” gender, on the left only females and males

5.



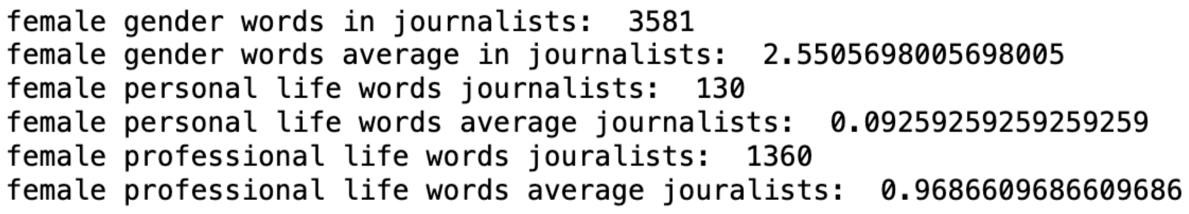
```
Stef_Blok 0.04656160458452722 M
Piet_Hein_Donner 0.03868194842406877 M
André_Rouvoet 0.03832378223495702 M
Halbe_Zijlstra 0.03832378223495702 M
Henk_Kamp 0.03832378223495702 M
Carola_Schouten 0.0379656160458452 7 F
Wouter_Bos 0.03796561604584527 M
Ronald_Plasterk 0.03796561604584527 M
Johan_Remkes 0.037607449856733526 M
Geert_Wilders 0.036891117478510024 M
Alexander_Pechtold 0.036891117478510024 M
Jetta_Klijnsma 0.03474212034383954 F
Edith_Schippers 0.034383954154727794 F
Fred_Teeven 0.03402578796561605 M
Jeanine_Hennis-Plasschaert 0.033309455587392546 F
Jan_Kees_de_Jager 0.0329512893982808 M
Job_Cohen 0.0329512893982808 M
Frans_Weekers 0.0329512893982808 M
Hanke_Bruins_Slot 0.0329512893982808 F
```



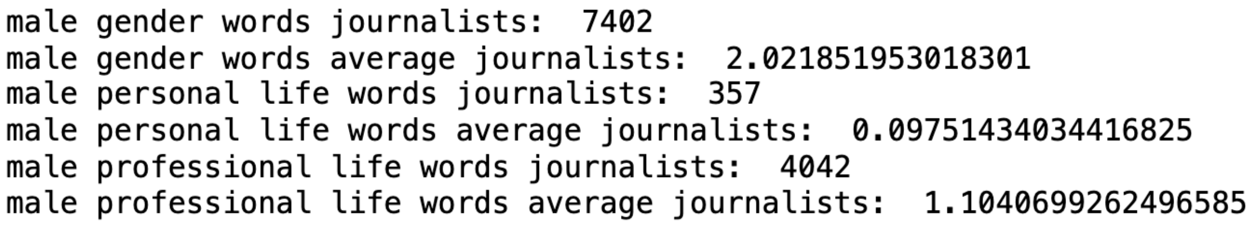
On the left the top people with their gender (F = female, M = male) on the right the network with the top 10 nodes by centrality highlighted

Journalists:

1. 4114 links between different genders and 15378 total links
2. The percentage of female that are local heroes is 25% while the male percentage is 23%
3.

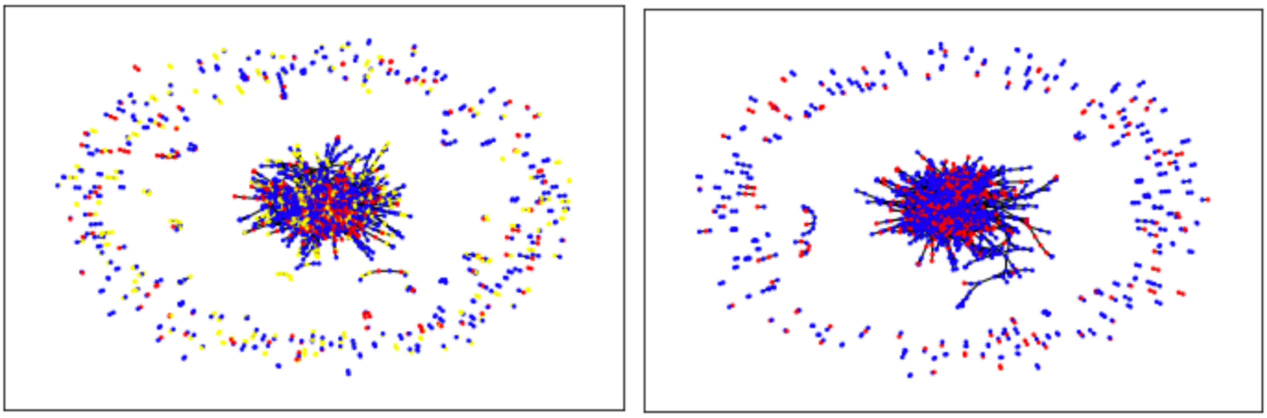


```
female gender words in journalists:  3581
female gender words average in journalists:  2.5505698005698005
female personal life words journalists:  130
female personal life words average journalists:  0.09259259259259259
female professional life words jouralists:  1360
female professional life words average jouralists:  0.9686609686609686
```



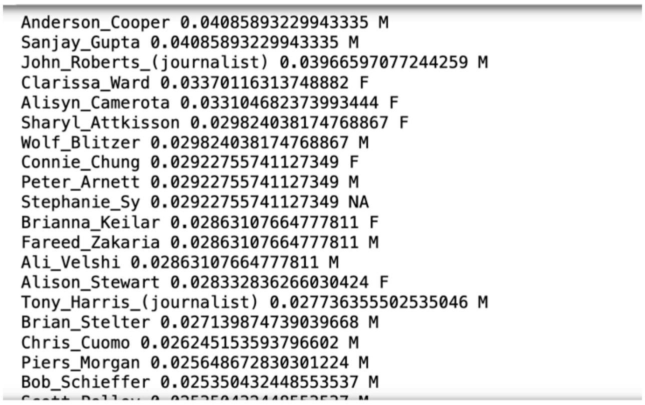
```
male gender words journalists:  7402
male gender words average journalists:  2.021851953018301
male personal life words journalists:  357
male personal life words average journalists:  0.09751434034416825
male professional life words journalists:  4042
male professional life words average journalists:  1.1040699262496585
```

4.

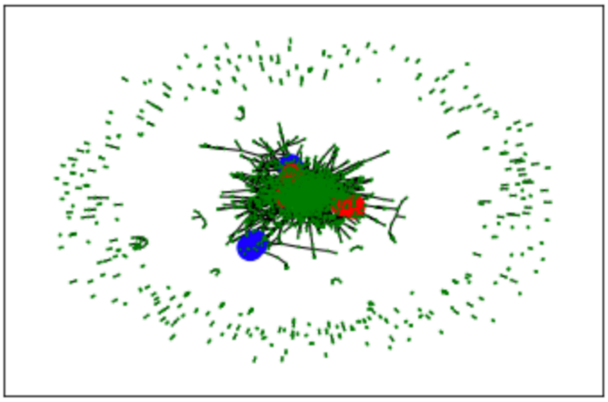


On the left the network including the “NA” gender, on the left only females and males

5.



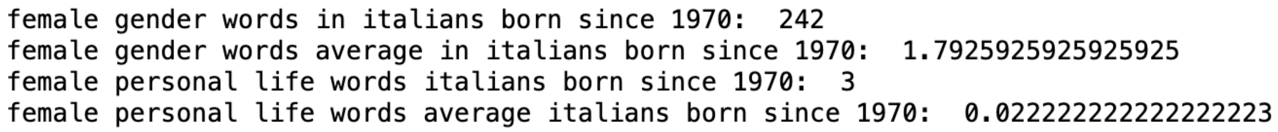
```
Anderson_Cooper 0.04085893229943335 M
Sanjay_Gupta 0.04085893229943335 M
John_Roberts_(journalist) 0.03966597077244259 M
Clarissa_Ward 0.03370116313748882 F
Alisyn_Camerota 0.033104682373993444 F
Sharyl_Attkisson 0.029824038174768867 F
Wolf_Blitzer 0.029824038174768867 M
Connie_Chung 0.02922755741127349 F
Peter_Arnett 0.02922755741127349 M
Stephanie_Sy 0.02922755741127349 NA
Brianna_Keilar 0.02863107664777811 F
Fareed_Zakaria 0.02863107664777811 M
Ali_Velshi 0.02863107664777811 M
Alison_Stewart 0.028332836266030424 F
Tony_Harris_(journalist) 0.027736355502535046 M
Brian_Stelter 0.027139874739039668 M
Chris_Cuomo 0.026245153593796602 M
Piers_Morgan 0.025648672830301224 M
Bob_Schieffer 0.025350432448553537 M
Scott_Pelley 0.025350432448553537 M
```



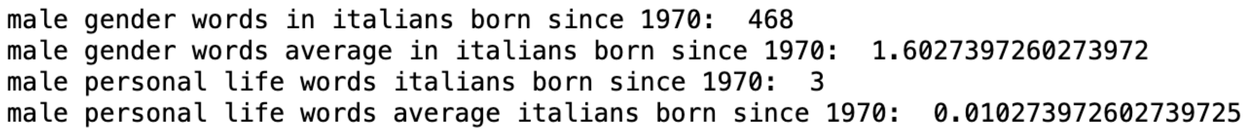
On the left the top people with their gender (F = female, M = male) on the right the network with the top 10 nodes by centrality highlighted

Italians born since 1970:

1. 509 links between different genders and 1958 total links
2. The percentage of female that are local heroes is 11.1% while the male percentage is 8.9%
3.

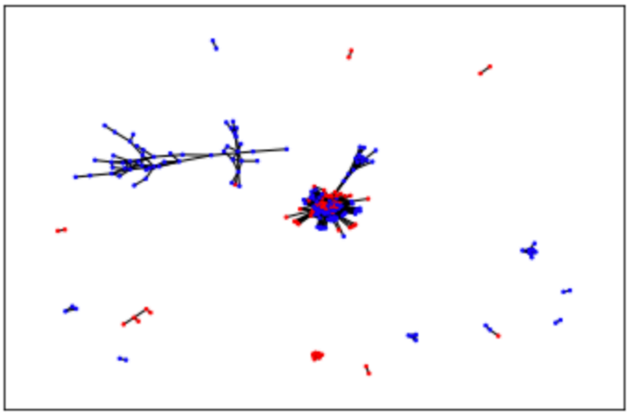


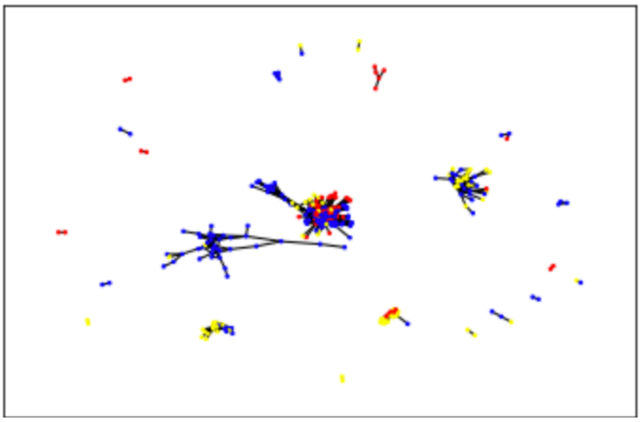
```
female gender words in italians born since 1970:  242
female gender words average in italians born since 1970:  1.7925925925925925
female personal life words italians born since 1970:  3
female personal life words average italians born since 1970:  0.022222222222222223
```



```
male gender words in italians born since 1970:  468
male gender words average in italians born since 1970:  1.6027397260273972
male personal life words italians born since 1970:  3
male personal life words average italians born since 1970:  0.010273972602739725
```

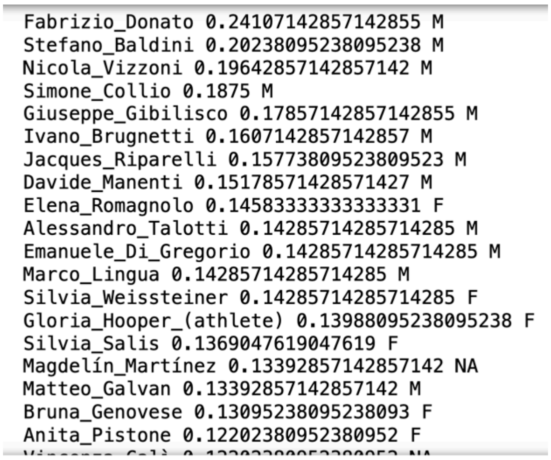
4.



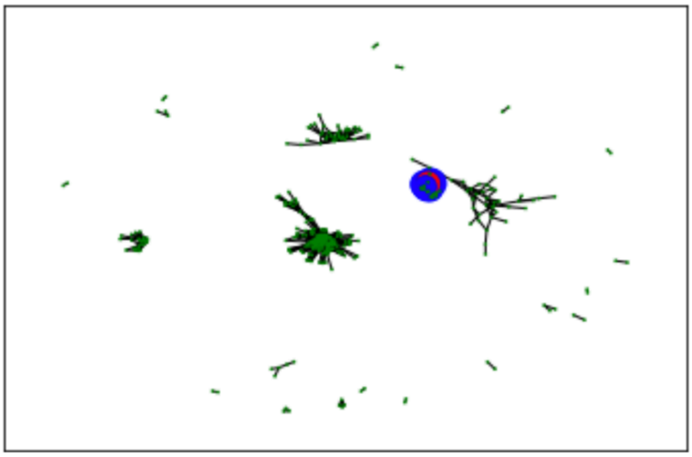


On the left the network including the “NA” gender, on the left only females and males

5.



```
Fabrizio_Donato 0.24107142857142855 M
Stefano_Baldini 0.20238095238095238 M
Nicola_Vizzoni 0.19642857142857142 M
Simone_Collio 0.1875 M
Giuseppe_Gibilisco 0.17857142857142855 M
Ivano_Brugnetti 0.1607142857142857 M
Jacques_Riparelli 0.15773809523809523 M
Davide_Manenti 0.15178571428571427 M
Elena_Romagnolo 0.14583333333333331 F
Alessandro_Talotti 0.14285714285714285 M
Emanuele_Di_Gregorio 0.14285714285714285 M
Marco_Lingua 0.14285714285714285 M
Silvia_Weissteiner 0.14285714285714285 F
Gloria_Hooper_(athlete) 0.13988095238095238 F
Silvia_Salis 0.1369047619047619 F
Magdelín_Martínez 0.13392857142857142 NA
Matteo_Galvan 0.13392857142857142 M
Bruna_Genovese 0.13095238095238093 F
Anita_Pistone 0.12202380952380952 F
Vincenzo_Calò 0.12202380952380952 NA
```



On the left the top people with their gender (F = female, M = male) on the right the network with the top 10 nodes by centrality highlighted

Italians born since 1950 (Italian Wikipedia):

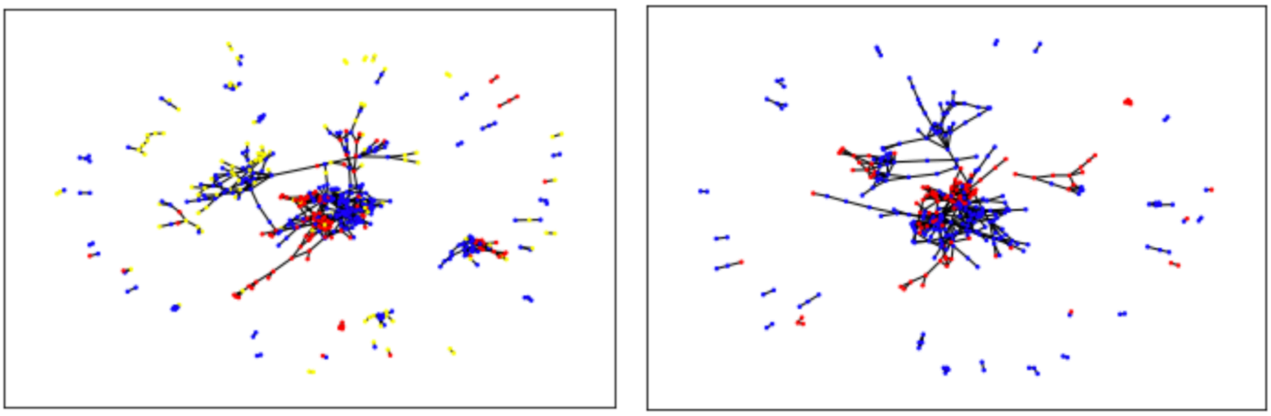
1. 180 links between different genders and 903 total links
2. The percentage of female that are local heroes is 13.3% while the male percentage is 5.9%
3.

```
female gender words in italians born since 1950:  2
female gender words average in italians born since 1950:  0.009478672985781991
female personal life words in italians born since 1950:  3
female personal life words average in italians born since 1950:  0.014218009478672985
```

```
male gender words in italians born since 1950:  14
male gender words average in italians born since 1950:  0.024390243902439025
male personal life words in italians born since 1950:  8
male personal life words average in italians born since 1950:  0.013937282229965157
```

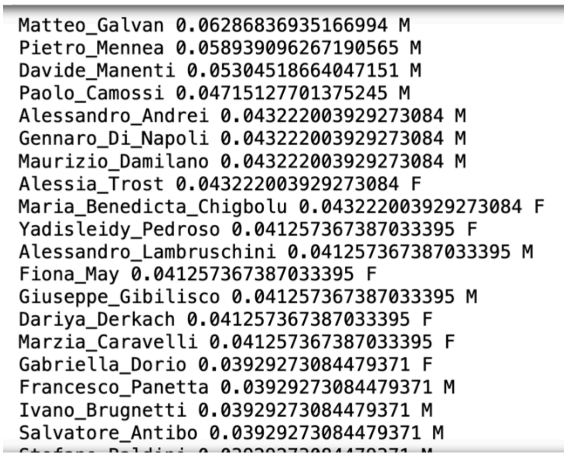
Here it can be seen how the Italian summaries are often short and don't contain many words in general, for this reason in this case this study cannot be considered for our intents.

4.

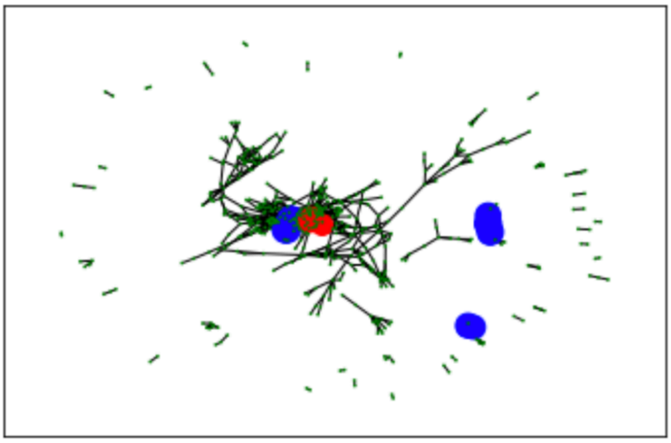


On the left the network including the "NA" gender, on the left only females and males

5.



```
Matteo_Galvan 0.06286836935166994 M
Pietro_Mennea 0.058939096267190565 M
Davide_Manenti 0.05304518664047151 M
Paolo_Camossi 0.04715127701375245 M
Alessandro_Andrei 0.043222003929273084 M
Gennaro_Di_Napoli 0.043222003929273084 M
Maurizio_Damilano 0.043222003929273084 M
Alessia_Trost 0.043222003929273084 F
Maria_Benedicta_Chigbolu 0.043222003929273084 F
Yadisleidy_Pedroso 0.041257367387033395 F
Alessandro_Lambruschini 0.041257367387033395 M
Fiona_May 0.041257367387033395 F
Giuseppe_Gibilisco 0.041257367387033395 M
Dariya_Derkach 0.041257367387033395 F
Marzia_Caravelli 0.041257367387033395 F
Gabriella_Dorio 0.03929273084479371 F
Francesco_Panetta 0.03929273084479371 M
Ivano_Brugnetti 0.03929273084479371 M
Salvatore_Antibo 0.03929273084479371 M
Stefano_Baldini 0.03929273084479371 M
```



On the left the top people with their gender (F = female, M = male) on the right the network with the top 10 nodes by centrality highlighted.

Conclusions:

Wrapping up we can say that there is a gender gap, in the way people are described in their summaries, from points 3 we can see that women on average often have more words related to their gender and fewer words related to their professional field. From points 1 we can say that there may actually be homophobia. On the other hand, in points 2 there is not always a gap between males and females in being local heroes, just as there is not always a big difference in the average amount of languages in which a biography is written.

As regards the study on occupations done on the Italian datasets the results seemed wrong, probably due to how the occupations are written in those datasets, for this reason the results were not reported, as well as for the dataset of Italian writers which is too small to be statistically interesting, however the code can be found below.

Moreover, despite the fact that it is generally males who are more central in the networks, on average the rank of females is often higher even if only slightly.