EURYKA

Reinventing Democracy in Europe: Youth Doing Politics in Times of Increasing Inequalities

Integrated Report on Social Media Analysis

(Deliverable 7.2)

Workpackage 10: Social Media Analysis

Workpackage Leading Institution: UOC

Submission due date: September 2019

Actual submission date: December 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727025

1. INTRODUCTION

The increasing use of social media in political debates is said to have contributed to new forms of participation in the public sphere. More generally, the development of technology, along with the immediacy and the ubiquity that the use of smartphones entails, accounts for new forms of communication that are increasingly integrated into people's daily routines. In the case of young people, this integration (although largely marked by the fact they are "digital natives") does not occur in a uniform or generalized manner, but rather unevenly. Therefore, the mere fact of being young does not guarantee greater access to online public debates. This part of the EURYKA project aims to analyze - among other things - how social inequalities are manifested in how young people use social media actively for political purposes.

The coordinated study (WP7) has been carried out in the nine countries participating in the project (France, Italy, the UK, Germany, Poland, Greece, Spain, Sweden and Switzerland). Specifically, the analysis has been centered on Twitter. EURYKA's project intended to analyse both Facebook and Twitter, but due to external conditionings only Twitter's data was available in the period of study. The plan was to use Netvizz[1] for Facebook analysis, which was the way to ensure a minimum of data retrieval since Facebook implemented restrictions on data access. We had to face the problem of Netvizz being no more active from August 21st. We could not find an alternative tool, so we could not collect any Facebook data.

The main objective of WP7 has been:

- To investigate young people's ways of doing politics online and the impact of inequalities on this by looking at the interactions taking place on Twitter.

The goal was to see how young people in these nine different countries participate and interact in the public debates around two important issues: the climate crisis and feminism. In order to make data retrieval operational, two case studies were selected: Twitter's debates on #ClimateStrike (a global movement, studied at the country/language area level) and the local campaigns or movements on feminism taking place in each country/language area.

This work package has faced two important challenges in relation to the project's development:

1) There is no direct access to the personal data (e.g. age, gender, geographical location) of Twitter and Facebook users.
2) Traditional tools of social media analysis do not deliver data that is representative of plurilingual realities. There is a bias of statistical inference tools towards dominant languages and groups.

[1] https://github.com/bernorieder/netvizz

2. METHODOLOGY

2.1. Data collection

For this study, data were collected from Twitter for the nine countries included in the study: France, Germany, Greece, Italy, Poland, Spain, Sweden, Switzerland, the United Kingdom. Additional data were collected by monitoring keywords at global level.

2.1.1. Keyword selection process

In this section, we specify the guidelines that were followed by each local partner to identify relevant keywords in their country (see Deliverable D7.1). Keywords were selected in collaboration with the partners of the project consortium from each country. In particular, they had to search for the most popular and relevant keywords and accounts related to *climate change ("fridaysforfuture")* and *feminism* topics in each country on Twitter. The guidelines given to the local partners were as follows:

1. *Open your browser in incognito mode/private browsing. Connect to the Twitter website and log into EURYKA's account (login and password will be sent by email to each partner). This will prevent the search from being done based on the algorithms generated by the researchers' personal accounts.*
2. *Go to the search bar. Type the search term of each topic:*
    a. *"fridaysforfuture" in English.*
    b. *the most representative translation/country-specific version of the word "feminism".*
3. *On the left of the screen, click on show search filters, then click on the advanced search below.*
4. *In "These hashtags", enter the keyword you had initially entered (e.g. "fridaysforfuture").*
5. *The teams of France, Switzerland and the United Kingdom will have to add location(s) in "Places". Please, try first the capital city, then secondary cities if the search does not yield enough results. In Switzerland, please use different cities corresponding to the different languages), then click search.*
6. *The teams of Germany, Sweden, Italy, Greece and Poland will select their language in "written in" and then click search.*
7. *Scroll through the first 100 tweets and look for relevant co-occurring hashtags (e.g. hashtags that are included in the tweets along with #fridaysforfuture and that are directly related to the movement). Select **at least 5** of these keywords for each topic. These keywords could include: a) hashtags b) Twitter accounts from **organisations** or **public profiles** (no personal accounts are allowed).*
8. *Fill in the Excel sheet with the relevant keywords that you found (note that in the tabs below there is one sheet for fridaysforfuture and another sheet for feminism).*

***Only if needed**, this task could be complemented with free online tools like Hashtagify[2]. You can use Hashtagify by writing the hashtag/keyword in the search bar and clicking search. This will give you - among other things - the popularity score, a cloud with the most popular related hashtags used along this hashtag, the language(s) in which it is used, as well as a world map with the countries in which the hashtag has been used the most.*

Potential problems

If a country partner cannot find enough results (minimum five keywords per topic):

- Scroll through the first 200 or 300 tweets, instead of the first 100, while running relevant hashtags in Hashtagify to see suggestions of additional related hashtags.

If a country partner finds it difficult to obtain enough results as far as the topic "feminism" is concerned:

- Repeat the process substituting the country-specific word for "feminism" by one (or more) of the relevant related hashtags that you found (for example, in Spain, we could start a new search with #feminista or #machismo instead of limiting ourselves to #feminismo).

2.1.2. Selected keywords by country

The table below presents the keywords that were finally selected for the data collection process of each country, plus the ones that were picked at a global level. For each country (as well as for global level) we report the hashtags selected for the two topics chosen for the project: climate change ("fridaysforfuture") and feminism.

We should note some exceptions with specific countries:

- **Switzerland**: Some relevant hashtags were identified by the local partners, however in the case of climate change they were either global (in English) or shared with the French, German or Italian communities, where most users are not from Switzerland, but from France, Germany or Italy respectively. For this reason, we did not track any specific hashtag on Climate Change for this country. For feminism, specific relevant hashtags from the Swiss feminist movement were identified (listed in the table below), but they were related to past events, and no longer active at the time of data collection.
- **United Kingdom**: Given the global scope of the English language, all the hashtags that were identified for the United Kingdom were not specific to the country, but were used at a global level, so we could not collect data for the United Kingdom separately.
- **Poland**: No specific hashtags were selected manually for Poland.
- **Greece**: Only a few hundred tweets could be retrieved with the identified hashtags.

[2] See https://hashtagify.me/

For these cases we could not collect data based on specific hashtags with the standard procedure, so we developed alternative solutions, detailed in the next section “Dataset description”, that allowed us to collate a dataset for each of the nine countries.

Country	Topic	Hashtags
Global	Climate Change	*#fridaysforfuture,#climatestrike,#schoolstrike4climate,#extinctionrebellion,#rebelforlife,#climateaction,#youthstrike4climate,#youthforclimate,#climateemergency,#youth4climate,#climatebreakdown,#gretathunberg,#climatejustice*
	Gender	*#feminist,#feminism,#metoo,#genderequality,#heforshe,#feminazi, #feminazis*
France	Climate Change	*#marchepourleclimat,#justiceclimatique,#grevemondialepourleclimat,#changementclimatique,#ilestencoretemps,#marcheclimat,#generationclimat,#grevepourleclimat,#printempsclimatique,*
	Gender	*#feminisme,#écoféminisme,#feministe,#sexisme,#meufpower,#actionféministe,#balancetonporc,#patriarcat*
Germany	Climate Change	*#klimagerechtigkeit,#klimaschutz,#klimakrise,#klimastreiks,#klimastreik,#endegelaende*
	Gender	*#feminismus,#sexismus,#gleichberechtigung,#frauen,#emanzipation,@marga_owski*
Greece	Climate Change	*#κλιματικήαλλαγη,#fff_greece,#ResilientAthens,#κλιματικήαλλαγή*
	Gender	*#βιασμός,#βιασμος,#Βιασμός,#βιασμός,#βιασμόσ,#τοξικηΑρρενωποτητα,#τοξική_αρρενωπότητα,#τοξικήαρρενωπότητα,#τοξικηαρρενωποτητα,#τοξικη_αρρενωποτητα,#μισογυνισμός,#μισογυνισμος,#Μισογυνισμός,#Μισογυνισμος,#γυναικοκτονία,#γυναικοκτονια,#γυναικοκτονια,#γυναίκες,#Γυναίκες,#γυναικες,#Γυναικες,#φεμινίστριες,#φεμινιστριες,#φεμινιστριεσ,#φεμνίστριεσ,#Φεμινίστριες,#φεμινιστικό_κίνημα,#φεμινιστικόκίνημα,#Φεμινιστικοκινημα,#Φεμινιστικο_κινημα,#φαλλοκρατία,#φαλλοκρατια,#φαλοκρατία,#φαλοκρατια,#σεξισμος,#σεξισμός,#Σεξισμός,#φεμινισμός,#φεμινισμος,#πατριαρχία,#πατριαρχια,#Πατριαρχια,#Πατριαρχία,#ενδοοικογενειακήβία,#ενδοοικογενειακηβια,#ενδοοικογενειακή_βία,#ενδοοικογενειακη_βια,#κουλτουραβιασμου,#κουλτουρα_βιασμου,#κουλτούραβιασμού,#κουλτούρα_βιασμού,#Κουλτουραβιασμου,#Κουλτουρα_βιασμου,#Κουλτούραβιασμού,#Κουλτούρα_βιασμού*
Italy	Climate Change	*#emergenzaclimatica,#scioperoperilclima,#agireora,#toccaanoi,#fridaysforfutureitalia*

	Gender	*#femminismo,#femminista,#donne,#maschilismo,#lottomarzo,#femminicidio,#nonunadimeno,#siamomarea*
Spain	Climate Change	*#emergenciaclimática,#medioambiente,#sostenibilidad,#acciónclimática,#leydecambioclimáticoya,#nohayplanetab,#horadelplaneta,#huelgaporelclima,#juventudxclima,#justiciaclimática,#cambioclimático,#marchaporelclima,#crisisclimática,#porelclima,#15mclimático,#laespañaquequieresesecologista,#huelgaclimática*
	Gender	*#sororidad,#feministas,#feminista,#machismo,#machismomata,#violenciamachista,#violenciasmachistas,#bastaya,#violenciadegénero,#genero,#patriarcado*
Sweden	Climate Change	*#klimatstrejk,#klimatkris,#klimat,#klimataktion,#fridaysforfuturesverige*
	Gender	*#fempol,#jämställdhet,#jämpol,@kvinnohistoria*[3]
Switzerland	Climate Change	-
	Gender	#2019GreveFeministe,#Frauenstreik19,#scioperofemminista,#frauenstreik2019,FrauenstreikCH,#14juin2019,#GrèveFéministe2019,#GreveFeministe2019,#feministischerstreik,#grevedesfemmes,#grevedesfemmes2019

Table 1. Keywords for the data collection process in each country and topic

2.1.3. Twitter data retrieval

Tweets from Twitter were retrieved using Kalium, a tool developed by Eurecat that allows one to efficiently and flexibly manage the tracking of social network data in real time (Napalkova *et al,* 2018). Kalium ensures robustness, scalability and flexibility when it comes to recovering and managing social network data. This system was used to retrieve information from the Twitter streaming API[4], monitoring the hashtags identified as relevant for each country.

The datasets include all the tweets posted between July 12th and September 30th, 2019 with at least one of the hashtags reported above. This period is especially relevant for the #ClimateStrike movement that organized massive global strikes and demonstrations between September 20th and 27th. In this regard, our dataset contains the process of the formation and growth this movement over two months, including the preparation and the celebration of the strikes and

[3] We collected data for @kvinnohistoria but then did not include them in the datasets, as we are including only data based on hashtags.
[4] See https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter

demonstrations in the last week of September 2019.

Figure 1. Example of a popular tweet collected for the Climate Change.

2.1.4. Estimation of demographic information

Data about Twitter users found in the tweets of the dataset (authors, mentioned, and retweeted) were then processed and enriched with additional demographic information, namely gender, age range and geographic location.

2.1.4.1. Gender and age range

The gender and age range of users were estimated using the state of the art library M3Inference[5]. The tool relies on a deep learning model trained on multilingual data to infer gender and age range of users, based on the user name, the short bio text and the profile picture of the user

[5] See https://github.com/euagendas/m3inference

(Wang et al, 2019). In addition, the tool also infers whether a user account is a personal account or it corresponds to an organization.

The tool returns the estimations as the probability of a user to belong to a given class (male/female, age range, organization/not organization). In order to only include the most accurate estimations, and filter out noisy data, we used a threshold of 0.9, as done by Wang et al (2019). In our process, we only assign demographic information to users whose probability of not being an organization is above 0.9. Then, among these users, we only assign male/female gender to the ones whose probability of being male/female according to the tool is above 0.9, and analogously we assign an age range (<30 or >=30) when the probability of belonging to the more likely age range is above 0.9.

In the case of gender inference, the distribution of probabilities is bimodal, with peaks close to 100% probability of being female or male, respectively. This is why even with a threshold of 0.9 we are able to assign a gender to over 50% of the prediction users.

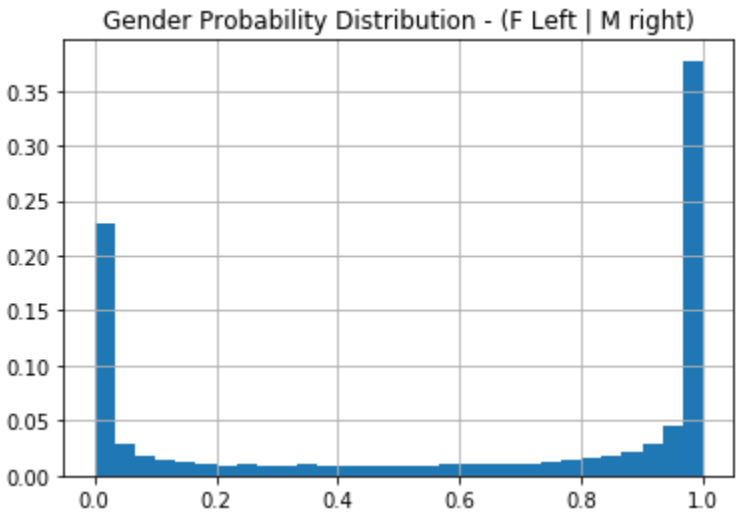

Figure 2. Distribution of the accuracy score of the gender inference for the users in our dataset. The X axis represents the probability of being male versus being female, the Y axis the corresponding proportion of users for each value of the probability: the two peaks indicate that for many users the probability is close to 100% (reliably identified as a man) or to 0% (reliably identified as a woman).

In the case of age, the situation is not as simple as for gender. The tool actually returns the probabilities for four age ranges: <18, 19-29, 30-39, >= 40. However, we observed that the accuracy of this prediction is much lower than for gender: for most users the probability of belonging to more classes is comparable, and for very few users there is a clear prediction for a determined age range. A threshold of 0.9 for each of the four age ranges in this case would leave the vast majority of the users unlabelled. Even with a lower threshold such as 0.7, only a small minority of users would overcome the threshold. We believe this is due to the fact that prediction

in the case of age is much harder for various reasons, including that the difference between age ranges is fuzzy: many users may lay on the border between two classes. Therefore we decided to merge the four classes provided by the tool into two larger classes: <30 and >=30. In this way, the probability scores get much higher and we are able to label a more consistent base of users with good accuracy. In other words, distinguishing between four age classes based on the profile pictures and short bios is hard, while distinguishing between two classes (below 30 or above 30) is easier and gives more accurate results. Figure 2 shows the distribution of the probabilities returned by the tool for the age prediction.

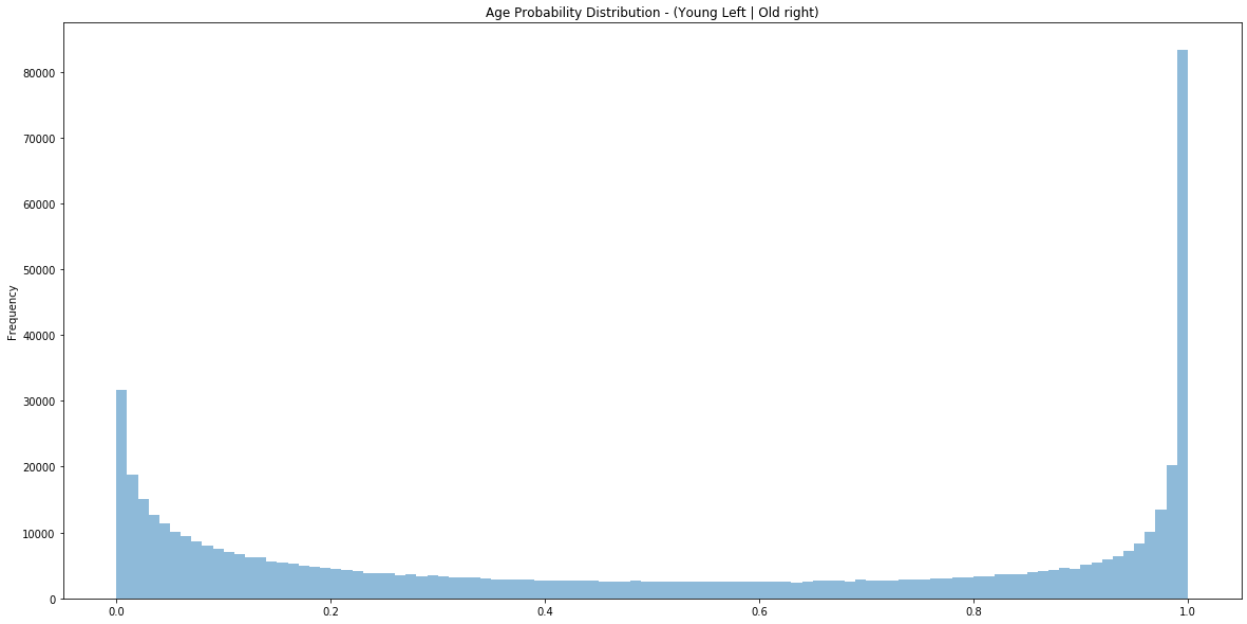

Figure 3. Distribution of the accuracy score of the age inference for the users in our dataset. The X axis represents the probability of being aged above 30 years old versus below 30 years old, the Y axis the corresponding proportion of users for each value of the probability: the two peaks indicate that for many users the probability is close to 100% (reliably identified as above 30) or to 0% (reliably identified as a below 30), although in this case the proportion of users with fuzzy probabilities is higher than for gender, indicating more uncertainty in the prediction.

We established the above thresholds to assign gender, age range and category (organization or not) for each user and included this information in the datasets. In addition, we included all the probabilities returned by the tool in the datasets in order to allow further analyses adopting different conventions, such as lower thresholds or keeping the four age ranges provided by the tool.

2.1.4.2. Geographic location

Each Twitter user is able to self-define its location. In some cases, the value of the user location field does not relate to a normative geographic location, e.g. “planet earth”, “BCN" (meaning

Barcelona), or even emoticons with country flags or any free text. However, most users filling this field reveal relevant geographic information. This information has no specific structure and language and no defined level of granularity, e.g., one user may write “Milan” or “Milano”, while others may write “Lombardia” or “Italy”.

To standardize all these values into countries and cities/regions, we relied on two open source libraries. First, we used the *Geocoder* tool[6] for mapping locations to geolocations on a map. In the case of values corresponding to a (latitude, longitude) pair, we used the *Reverse Geocode* tool[7] to identify the country and the city/region corresponding to that geolocation.

2.2. Dataset description

We have created 18 datasets, corresponding to our two topics (ClimateStrike and Feminism) for each of the nine countries. For each dataset we have one user file, with anonymized information about each user, and two interaction network files, with the networks of retweets and mentions, respectively.

In the following we describe the information included in the datasets, and the strategy we designed to collate the dataset for each country. For different reasons it was not possible to retrieve data for all countries based on keyword search, so for several anomalous countries we adopted different strategies.

2.2.1. User information

For each user, the datasets contain the following fields extracted directly from the Twitter API:

- ***user_followers_count_START***: number of followers at the beginning of the observation period (i.e., as of July 12th, 2019)
- ***user_followers_count_END****:* number of followers at the end of the observation period (i.e., as of September 30th, 2019)
- ***user_statuses_count_START****:* overall number of tweets posted by the user, at the beginning of the observation period (i.e., as of July 12th, 2019)
- ***user_statuses_count_END****:* overall number of tweets posted by the user at the end of the observation period (i.e., as of September 30th, 2019)
- ***user_created_at***: date of creation of the user account
- ***user_location***: user location reported by the user in his profile

The datasets further contain for each user additional fields representing demographic information inferred with other tools:

- ***tweets_count***: number of tweets posted by the user within the dataset

[6] https://geocoder.readthedocs.io
[7] https://pypi.org/project/reverse-geocode/

- ***is-org***: org or not org account estimated with m3inference
- ***gender***: gender estimated with m3inference
- ***age***: age range (either <30 or >= 30) estimated with m3inference
- ***city***: city corresponding to the user location
- ***country***: country code corresponding to the user location
- ***country_code***: code of the country corresponding to the user location (ISO 3166-1 alpha-2 standard)

Finally, the datasets include for each user metrics of centrality in the interaction networks. Such metrics are illustrated and detailed below in Section “Node metrics”.

2.2.2. Interaction networks

For each dataset, we created two networks: one based on retweets and the other based on mentions. Each interaction (each retweet, or mention) corresponds to an edge (connection) between two nodes that represent two users, and in which, according to an established convention, an incoming connection to a node represents received attention:

- **Retweet network**: if user A writes a tweet and user B retweets this tweet, a directed edge from B to A is generated in the retweet network.

- **Mention network**: if a user user A posts a tweet mentioning user B, a directed edge is generated in the mention network from A to B.

The networks are weighted, i.e. they include repeated edges (one link from A to B for each time A retweets/mentions B in a tweet). Most network metrics do not account for repeated edges, as we will see in the next section.

2.2.3. Datasets by country

For each of the two topics in each of the nine countries, we built a dataset with all tweets including the corresponding hashtags. This was the general rule for most countries except for some specific countries that required additional/different rules:

- **Spain:** For each of the two topics, we included all tweets including the specific hashtags selected for Spain (files with “spain” prefix). In this case, given the high presence of users from Latin America, we then filtered the files and created a filtered version including only edges in which at least one of the two users is located in Spain, and all the users involved in these edges.

- **Switzerland:** As explained in the Data collection section, in the case of Switzerland it was not possible to collect data based on specific hashtags, either because they were not specific for Switzerland, or because they had no or very little data. Instead, we filtered the two global datasets, and kept only the edges between two users when at least one of the

two users has a location in Switzerland. We then added in the users' files all users appearing in either of these networks.

- **United Kingdom:** As for Switzerland, we filtered the two global datasets and kept only the edges between two users when at least one of two users has a location in the United Kingdom. We added in the users' files all users appearing in either of these networks.
- **Poland:** For Poland no specific hashtags were monitored, however we were able to retrieve data from the global dataset. In this way we created two datasets for each topic: one including tweets in Polish, and one including tweets involving at least one user from Poland, as for Switzerland and the United Kingdom. We then merged the datasets collected in these two ways in one single dataset for each topic.
- **Greece:** For each of the two topics, we included all tweets including the specific hashtags selected for Greece. Given the very low amount of data retrieved with this method, we also included tweets in Greek language from the global datasets. We furthermore collected all tweets including users with user location in Greece, like in the case of the United Kingdom and Switzerland. Like in the case of Poland, we then merged the datasets collected in these ways in one single dataset for each topic.

2.3. Data analysis methods

2.3.1. Structural network metrics

In order to characterize social interactions in the different datasets, we computed a few structural metrics for each network. In the following we explain the meaning of each metric:

- **Number of nodes**: number of users having at least one interaction (retweet or mention) within the given dataset.
- **Number of edges**: number of connections (excluding duplicate connections, i.e. if user A retweeted user B 10 times, this will only count as one connection).
- **Clustering coefficient**: also known as transitivity, it represents the proportion of closed triangles in the network, over all the possible triangles (Watts and Strogatz, 1998), e.g., the proportion of cases in which if A is connected to B and B is connected to C, then C is connected to A.
- **Giant component**: the absolute and relative number of nodes in the largest connected component (weakly connected, i.e. we do not consider the direction of edges for computing this metric). It represents the size of the largest group of nodes that are all connected to one another through some path.
- **Average path distance**: the average distance between two nodes of the network graph. This property implies that all nodes are interconnected through a small number of steps

from one to another. It is important because the shorter the average distance, the faster is the spread of information over a network.

- **Reciprocity**: proportion of reciprocal connections. I.e., proportion of cases in which user A retweets/mentions user B, and B in turn retweets/mentions user A.
- **Density**: the ratio between the number of edges and the number of possible edges, given the number of nodes (Wasserman and Faust, 1994).
- **Gini indegree**: Gini coefficient of the in-degree distribution: it shows the skewness of the distribution of attention received by the users in the graph.
- **Gini outdegree**: Gini coefficient of the out-degree distribution. it shows the skewness of the distribution of attention given by the users in the graph to other users.

2.3.2. Node metrics

We computed several established metrics that quantify centrality of nodes according to different criteria. We included in the datasets the centrality of each node in both networks (retweets and mentions), as well as the ranking of the node for every individual metric. The metrics are the following ones:

- **In-degree**: number of incoming connections, i.e., number of distinct nodes that have retweeted/mentioned the user
- **Out-degree**: number of outgoing connections, i.e., number of distinct nodes that the user has retweeted/mentioned
- **Closeness**: indicates how close a node is to the other nodes in the network, i.e. how easy it is (technically, how many steps it takes, or through how many other nodes one should pass) from a node to reach the other nodes, on average. This is measured as the reciprocal of the sum of the distances between a given node and all the other nodes in the network (the shorter the distance, the higher the closeness centrality).
- **Coreness** (or **k-index**) measures how much a user is in the core of a network. Technically, a node has k-index k if it is connected to at least k other nodes which also have k-index at least k, as shown in the figure below. So, a high k-index means that a node is part of an inner core of the network, made of a group of nodes that have many connections with each other.
- **Pagerank**: Pagerank is like in-degree, but connections from relevant nodes are given a higher weight. Intuitively, the pagerank of a node represents the probability that, following a random path in the network, one will reach that node (Page, 1999). The algorithm is an iterative process, as the Pagerank of a node depends on the Pagerank values of the nodes that link to it, however there are fast algorithms to compute it.

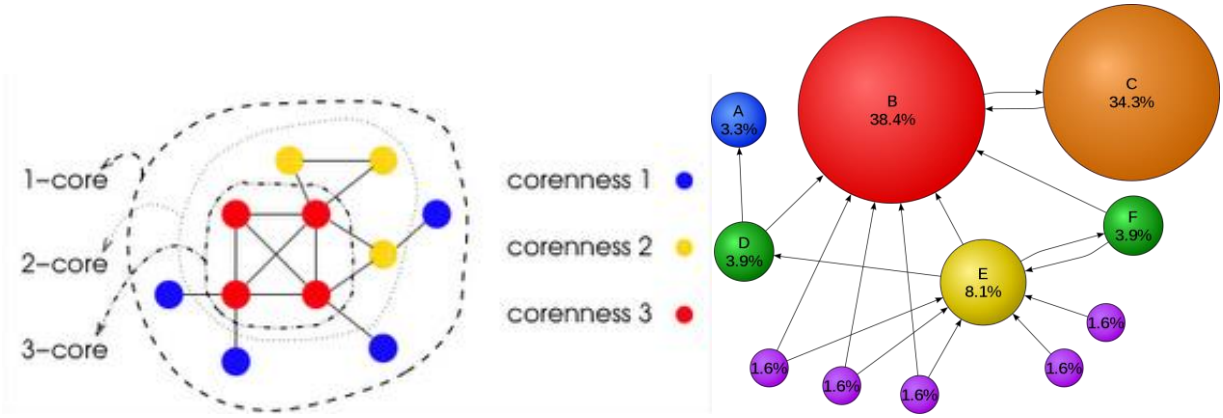

Figure 4. K-core decomposition process (left) and PageRank process (right).
Sources: Alvarez-Hamelin et al. (2005) and Wikipedia.

2.3.3. Homophily metrics

Homophily is a well-known phenomenon in network science and can be expressed as the "tendency of a group of users to link one each other". In our case, where users can present different characteristics, a subgroup of users can be less or more homophilic, driven by the exhibited attribute. We define homophily as a score which evaluates the observed intra-group edges distribution with respect to a random configuration. In formula:

$$h_i = \frac{|E_{ii}|}{|E_{i.}|} - \frac{n_i - 1}{N - 1}$$

The second term normalises the score, expressing the tendency of a node to connect to a peer in absence of homophily, i.e. the case in which the probability is given by the nodes distribution. The score is expressed within the range [−1,1] and is positive when the tendency of connecting to nodes of the same group is stronger than expected, negative when nodes are heterophilic, and equal to zero when the nodes behave as expected in a random configuration.

In order to compare how a subpopulation within the network connects one each other, with respect to the rest of the graph, we use the connectance, a metric which measures the average probability to be connected of two nodes in the same group (Park, 2007).

$$p_{ii} = \frac{|E_{ii}|}{(n_i - 1)n_i}$$

This probability can be useful for comparing the connectivity levels between two subgroups,

through the log odds.

$$log(\frac{p_{ii}/(1 - p_{ii})}{p_{jj}/(1 - jj)})$$

A positive value means the group ***i*** presents a strong tendency to connect more often than group ***j***, while a negative value the opposite situation.

2.3.4. Inequality metrics

For assessing inequality in our datasets, we compared the distributions of activity (number of tweets) and received attention (in-degree in the retweet network) between different groups of users.

Namely, we compare the distributions by gender (male users vs female users) and age (users below 30 years old vs users above 30 years old). Furthermore, to assess inequalities by gender within young users, which are the object of analysis of the project, we compared the distribution by gender among users below 30.

3. RESULTS

In this section, we report the results obtained for each country and for each topic, including macroscopic measures of the interaction networks and the analysis of inequalities and of homophily in these networks, illustrating the results with visualizations highlighting demographic characteristics of the users.

3.1. Cross-country comparison

We start by presenting an overview of the comparison between countries according to different aspects.

3.1.1. Structural network metrics

In the following, we present the macroscopic properties of the networks of interaction, comparing the networks in different countries. The four tables represent interactions based on retweets and mentions, for the two selected topics, Climate Change and Feminism.

The first two columns account for the size of the networks in number of nodes and connections (edges). When looking at the size of the networks it should be remembered that the differences

may be partly due to the heterogeneity of the scenario under analysis, stemming possibly from different usages of local versus global hashtags and languages, and from the different criteria used for data collection to adapt to the linguistic and geographic boundaries for each country.

The big size of the networks created for the UK may be interpreted in light of high levels of Twitter activity in the country and of engagement in social movements, together with the fact that the language of the country corresponds to the global language of the movements, so the criterion used for generating the UK dataset, i.e. tracking global hashtags and then considering conversations involving at least one user located in the UK, has resulted in big datasets in both cases, and especially in the case of climate strike movement. The Swiss networks, that were created in the same way, are also of considerable size in comparison with the limited population of the country, highlighting a high presence on Twitter and high participation with global hashtags also in Switzerland, especially in the climate strike movement.

Smaller networks were obtained with this method in the Polish and Greek cases, suggesting a lower online participation, although it is possible that we missed some conversations with local hashtags we were not tracking. In the case of Spain, the size is smaller than the overall amount of data collected for the language because, given the high presence of users from Latin America, data retrieved for Spanish hashtags were also filtered keeping only interactions were at least one user was located in Spain; in this way, the resulting networks obtained are relatively smaller than the ones created for other languages, as Spanish users with no self-reported location have also been discarded. For France, Germany, Italy and Sweden, we obtained networks of considerable sizes based on hashtags in the local languages; it should be reminded that in these cases the networks do not coincide exactly with the countries, but with the languages, so for example the German networks include also data for Austria, etc.

Looking at the macroscopic properties of the networks, we see that in most cases a vast majority of the nodes are all connected with each other through some path in the so-called giant component; more exactly the values, reported both in absolute and relative terms, are computed for the largest weakly connected component, i.e. links in any directions are considered (if considering the largest strongly connected component, the values would be much lower, as one would require to have paths in both directions between two nodes in order to consider them to be connected). We observe that in the cases of large and dense networks, the giant component typically encompasses over 90%. Only in few cases of smaller networks we have a fragmented structure with less than 75% of the nodes belonging to the giant component; this is the case for many mention networks on Feminism, where for several countries the interactions are more sparse, and we do not observe the emergence of a big core of users all connected to each other.

The values of clustering coefficient account for the presence of triangles in the network, i.e. transitive connections, and indicate the higher or lower presence of cohesive community structure. Higher values are typical of human social networks and have been related to higher network resilience: removing one or a few nodes has a lower impact as in a cohesive structure many alternative paths may typically exist. Low values are typical for cases where the network is lacking a community structure, either because it is centralized around some central node, or

because it is made of sporadic interactions; this is the case for specially low values encountered in sparse networks such as the mentions networks for countries like Greece or Switzerland. Higher values of the clustering coefficient indicate the presence of a community structure, where users do not only give attention to some central node, but there is a group of active users that also interact with one another through retweets and mentions. We find especially high values of clustering coefficient in the networks of the climate strike movement, especially for countries such as Sweden, Italy and Germany, indicating that in these cases we have a larger community of engaged users actively involved in the conversation, interacting with one another and creating a cohesive structure.

The values of reciprocity represent the proportion of interactions that are reciprocated. The values tend to be low, especially in larger networks and in retweet networks, as these interactions are less likely to be reciprocated; for example, many users only participate in an online campaign retweeting other users, without creating original content, so connections are often unidirectional.

For analogous reasons also the Gini coefficient of the in-degree distributions tends to be higher in retweet networks: retweet networks tend to be larger, less reciprocal, and more centralized, with many users retweeting a few central nodes. We see the highest value of the Gini coefficient for Swedish retweet network on the climate strike movement, which seems to point out the particularly strong influence of a specific user, Greta Thunberg, in her own country and language community, on the social movement she inspired.

The Gini coefficient of the in-degree distribution measures the inequality of the distribution of attention received by users and can be seen as a proxy for the concentration of influence in each network; in the next sections we will take a deeper look into these inequalities, visualizing the distributions of in-degree by country and accounting for demographic factors such as age and gender.

3.1.1.1. Climate Change

Country	Nodes	Edges	Clustering coeff	Giant comp.	Giant comp. %	avg distance	reciprocity	density	Gini indegree	Gini outdegree
France	20596	36857	0.112	19504	94.70%	3.94	0.001	0.000087	0.986	0.447
Germany	128278	452500	0.163	124368	96.95%	6.23	0.005	0.000027	0.981	0.656
Greece	17397	20980	0.045	16187	93.04%	2.71	0.002	0.000069	0.952	0.450
Italy	56235	126326	0.122	54107	96.22%	7.33	0.003	0.000040	0.985	0.517
Poland	23195	28766	0.087	21590	93.08%	6.24	0.002	0.000053	0.971	0.382

Spain	23327	27648	0.076	19231	82.44%	6.03	0.004	0.000051	0.962	0.315
Sweden	49990	67668	0.233	48723	97.47%	6.62	0.003	0.000027	0.996	0.290
Switzerland	38845	55076	0.069	36388	93.67%	7.48	0.005	0.000037	0.940	0.539
United Kingdom	538724	1095573	0.113	525658	97.57%	6.50	0.005	0.000004	0.987	0.514

Table 2. **Climate change retweet networks.** Structural metrics of the network of retweets for each country dataset about climate change (Nodes: number of nodes, Edges: number of edges (connections), Clustering coeff: global clustering coefficient, Giant comp.: number of nodes in the largest connected component, Giant comp. %: percentage of nodes in the largest connected component, avg distance: average path distance between two nodes, reciprocity: proportion of reciprocal connections, density: density of connections in the network, Gini indegree: Gini coefficient of the indegree distribution, Gini outdegree: Gini coefficient of the outdegree distribution).

Country	Nodes	Edges	Clustering coeff	Giant comp.	Giant comp. %	avg distance	reciprocity	density	Gini indegree	Gini outdegree
France	1832	1595	0.055	829	45.3%	1.16	0.001	0.000475	0.666	0.663
Germany	41887	130868	0.357	38815	92.7%	6.62	0.034	0.000075	0.862	0.685
Greece	2281	2778	0.184	1791	78.5%	1.68	0.050	0.000534	0.684	0.705
Italy	13953	26484	0.369	12631	90.5%	11.62	0.028	0.000136	0.853	0.628
Poland	6003	11172	0.419	5452	90.8%	7.24	0.053	0.000310	0.823	0.646
Spain	10093	10368	0.040	7559	74.9%	2.37	0.001	0.000102	0.348	0.925
Sweden	11006	26087	0.457	10405	94.5%	6.11	0.058	0.000215	0.873	0.632
Switzerland	15144	20252	0.131	14066	92.9%	10.79	0.043	0.000088	0.871	0.574
United Kingdom	153629	300423	0.246	144056	93.8%	9.33	0.043	0.000013	0.889	0.619

Table 3. Climate change mention networks. Structural metrics of the network of mentions for each country dataset about climate change (Nodes: number of nodes, Edges: number of edges (connections), Clustering coeff: global clustering coefficient, Giant comp.: number of nodes in the largest connected component, Giant comp. %: percentage of nodes in the largest connected component, avg distance: average path distance between two nodes, reciprocity: proportion of reciprocal connections, density: density of connections in the network, Gini indegree: Gini coefficient of the indegree distribution, Gini outdegree: Gini coefficient of the outdegree distribution).

3.1.1.2. Feminism

Country	Nodes	Edges	Clustering coeff	Giant comp.	Giant comp. %	avg distance	reciprocity	density	Gini indegree	Gini outdegree
France	30667	56480	0.117	29090	94.86%	4.39	0.0016	0.000060	0.985	0.461
Germany	13521	19148	0.074	11607	85.84%	4.49	0.0024	0.000105	0.960	0.412
Greece	1431	1081	0.022	146	10.20%	1.35	0.0019	0.000528	0.756	0.505
Italy	8715	11621	0.114	7964	91.38%	4.53	0.0015	0.000153	0.983	0.318
Poland	3113	2816	0.016	970	31.16%	1.02	0.0014	0.000291	0.961	0.241
Spain	32518	44960	0.165	30178	92.80%	5.88	0.0046	0.000043	0.986	0.325
Sweden	1417	1744	0.073	1102	77.77%	1.94	0.0046	0.000869	0.937	0.382
Switzerland	5318	4894	0.014	3183	59.85%	3.60	0.0058	0.000173	0.927	0.330
United Kingdom	56116	61647	0.085	43579	77.66%	11.85	0.0076	0.000020	0.945	0.347

Table 4. **Feminism retweet networks.**Structural metrics of the network of mentions for each country dataset about feminism (Nodes: number of nodes, Edges: number of edges (connections), Clustering coeff: global clustering coefficient, Giant comp.: number of nodes in the largest connected component, Giant comp. %: percentage of nodes in the largest connected component, avg distance: average path distance between two nodes, reciprocity: proportion of reciprocal connections, density: density of connections in the network, Gini indegree: Gini coefficient of the indegree distribution, Gini outdegree: Gini coefficient of the outdegree distribution).

Country	Nodes	Edges	Clustering coeff	Giant comp.	Giant comp. %	avg distance	reciprocity	density	Gini indegree	Gini outdegree
France	3116	2842	0.055	1560	50.06%	1.33	0.0014	0.000293	0.666	0.678
Germany	5903	6109	0.095	3530	59.80%	2.18	0.0106	0.000175	0.612	0.739
Greece	550	388	0.000	35	6.36%	1.91	0.0052	0.001285	0.415	0.752
Italy	1885	1736	0.077	863	45.78%	1.43	0.0035	0.000489	0.561	0.777
Poland	1271	1080	0.044	239	18.80%	1.84	0.0056	0.000669	0.353	0.860
Spain	3276	3248	0.030	1877	57.30%	1.67	0.0034	0.000303	0.652	0.729
Sweden	1290	1384	0.175	876	67.91%	1.46	0.0290	0.000832	0.605	0.730
Switzerland	1729	1450	0.039	475	27.47%	1.88	0.0055	0.000485	0.500	0.785
United Kingdom	23074	22102	0.064	12395	53.72%	3.81	0.0110	0.000042	0.580	0.764

Table 5. **Feminism mention networks.**Structural metrics of the network of mentions for each country dataset about feminism (Nodes: number of nodes, Edges: number of edges (connections), Clustering coeff: global clustering coefficient, Giant comp.: number of nodes in the largest connected component, Giant comp. %: percentage of nodes in the largest connected component, avg distance: average path distance between two nodes, reciprocity: proportion of reciprocal connections, density: density of connections in the network, Gini indegree: Gini coefficient of the indegree distribution, Gini outdegree: Gini coefficient of the outdegree distribution).

3.1.2. Inequality by country

In the previous section we have seen the Gini coefficient of in-degree and out-degree distributions for each country, which we used to quantify the inequality of the attention received and given by users, respectively. As discussed in the “Data analysis methods” section, in our scenario in-degree is more relevant than out-degree, as the attention received from other users can be considered as a trustworthy indicator of influence in the network. On the contrary, out-degree is not especially relevant as with this metric it is easy to achieve a high centrality by just mentioning or retweeting many users; therefore, having a high out-degree does not necessarily imply being influential in the network.

In this section we focus on in-degree as a proxy for influence in the network, and look at the distributions of centrality by topic and by country.

Before looking at influence as measured by in-degree, we take a look at the distribution of activity, i.e. of the number of tweets posted by user. Like out-degree, this is not an indicator of influence, but we use it to quantify engagement, and look at how skewed is the distribution of activity among users in different countries.

In all the graphics of distributions, we plot on the X axis the different values of the metric of interest (activity, i.e. number of tweets, or in-degree centrality, i.e. number of incoming connections from distinct users), and on the Y axis the number of users having that value for that metric. This allows us to observe in which way a variable is distributed among users, and to what extent it is concentrated, with many users having low values of activity/centrality, and few users having higher levels of online engagement or influence in the movements.

3.1.2.1. Activity inequality

The figure shows the distribution of activity by user: the typical heavy tailed distribution indicates that the majority of the users posted only one or few tweets in the dataset, and a few users posted a high number of tweets (up to thousands of tweets in the case of the UK dataset).

We observe that the distributions tend to have a similar shape, only shifted by number of users, with the UK having the larger base of users and levels of activity reached. We observe some differences in the shape of the distribution for some countries, especially Switzerland and Poland in the climate strike movement, where the lines generated by the distributions tend to be more curved, indicating in proportion a lower presence of less active users, and a higher presence of engaged users posting a high amount of messages.

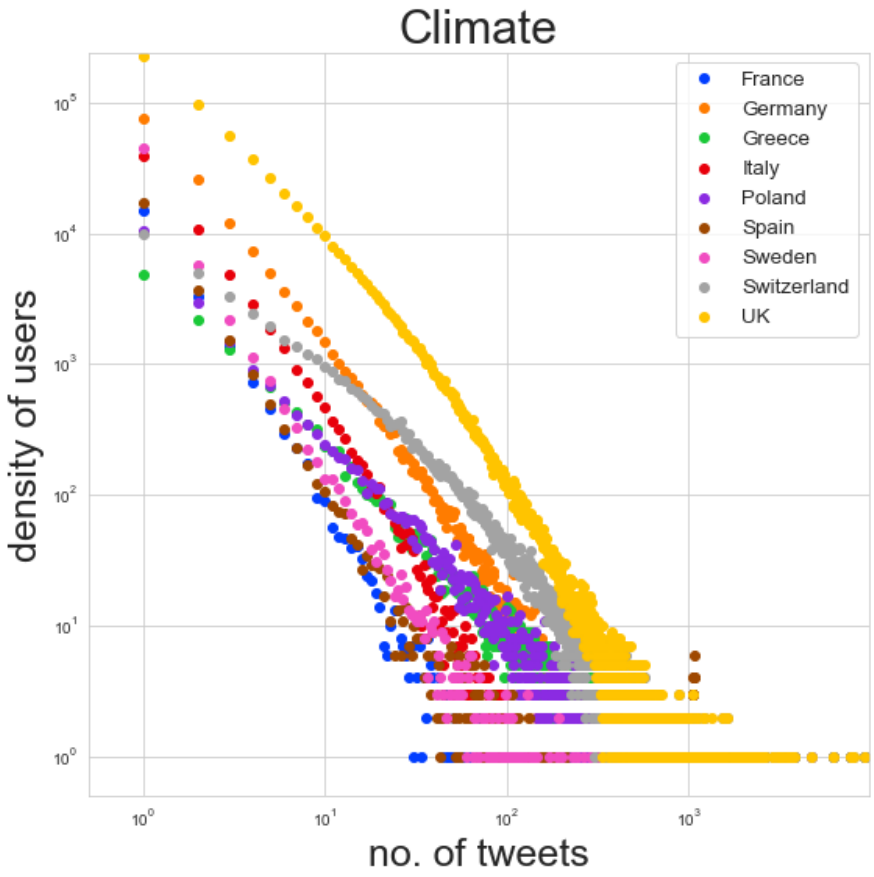

Figure 5. Distribution of users by number of tweets in the Climate change dataset for each country.

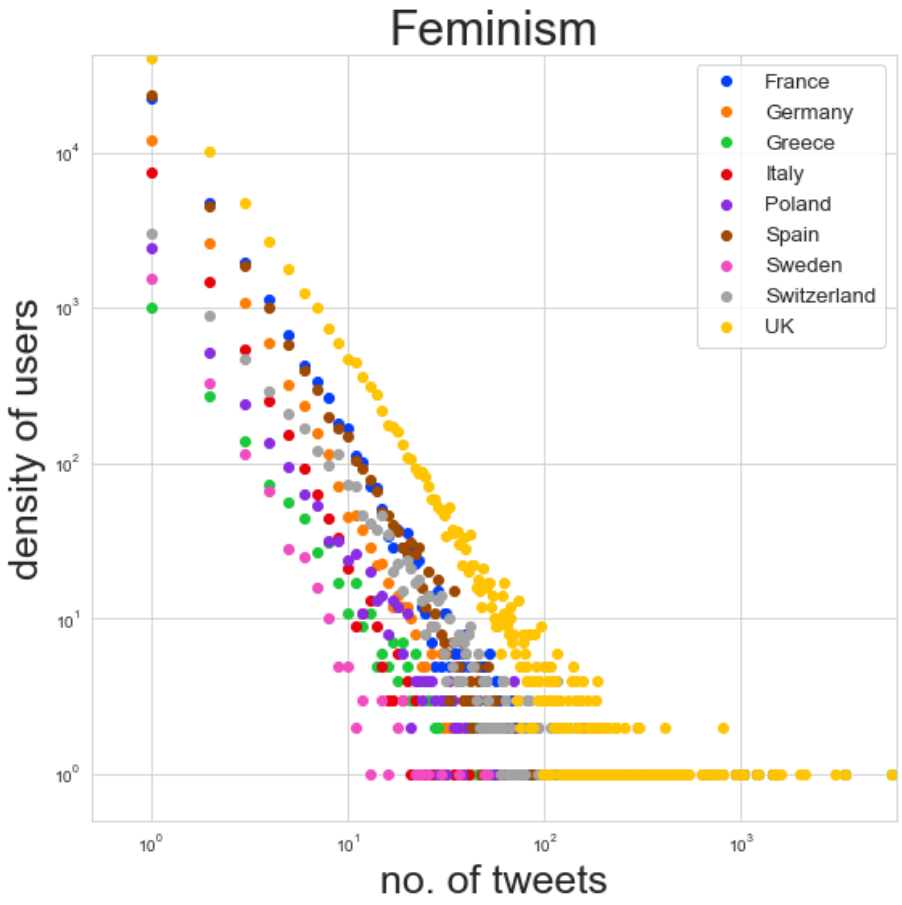

Figure 6. Distributions of users by number of tweets in the Feminism dataset for each country.

3.1.2.2. Centrality inequality

We now look at the distributions of in-degree centrality in the networks of retweets and mentions, comparing results by country.

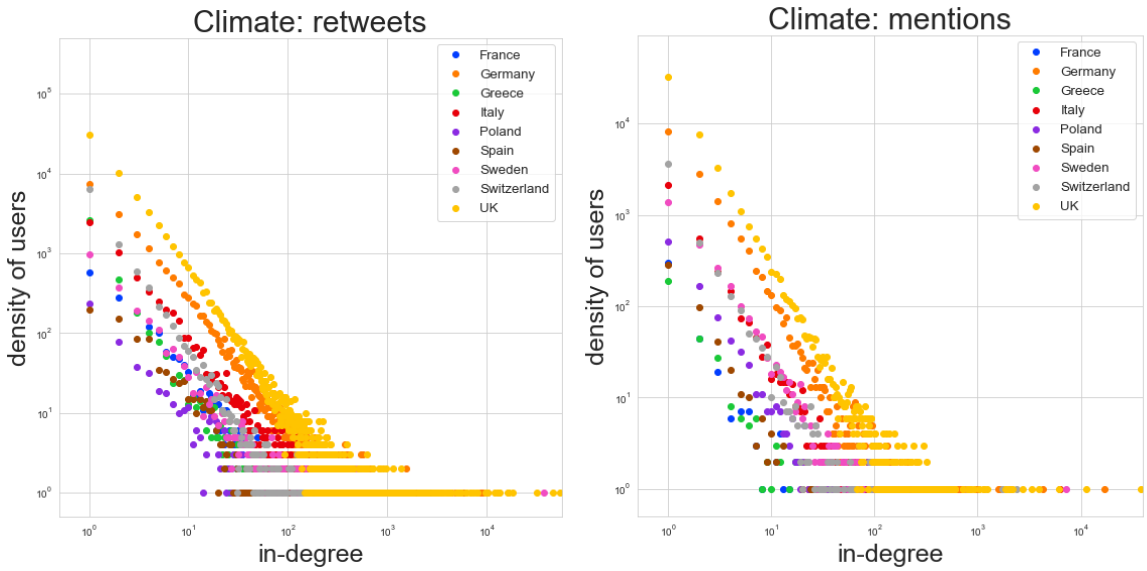

Figure 8. Distributions of users by indegree in the Climate change dataset for each country, in the networks based on retweets (left) and mentions (right).

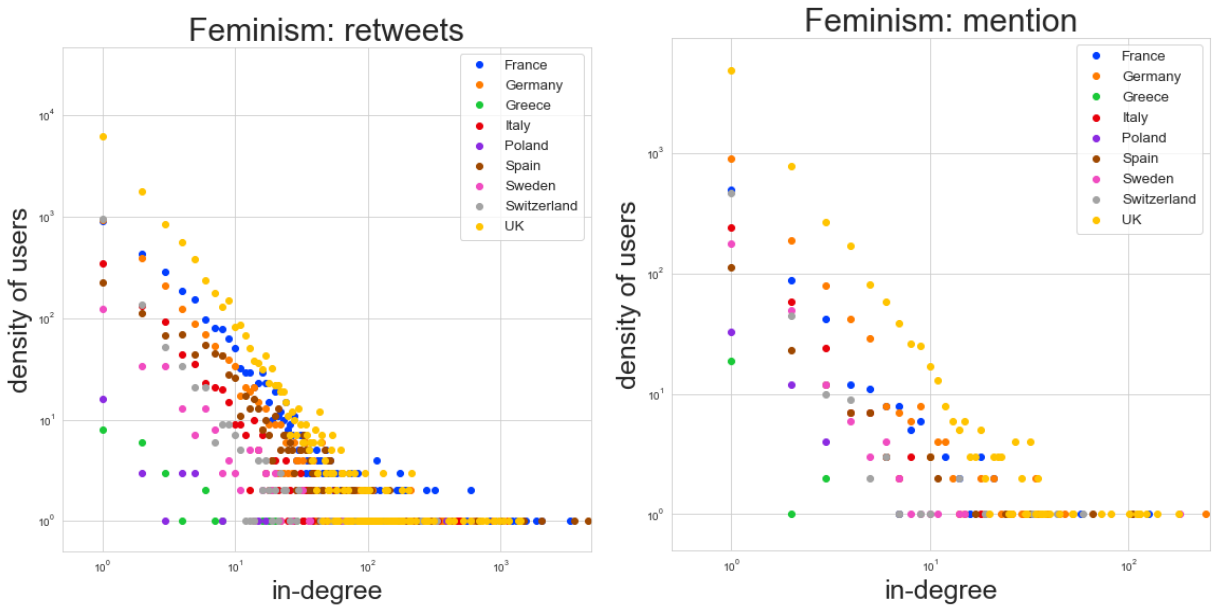

Figure 9:. Distributions of users by indegree in the Feminism dataset for each country, in the networks based on retweets (left) and mentions (right).

3.2. Homophily analysis results

In the following, we present an analysis of the demographic composition of the users in each network, and of their preference for interacting with other users with common characteristics. In particular, we focus our analysis on gender and age.

In each table, we report the proportion of users identified within each class (male/female, below/above 30 years old); it should be noticed that the percentages do not sum up to 100%, as the normalization is done among all users, i.e. including also users for which we could not infer their demographic characteristics. Actually, in the case of age, it can be observed that only for a minor fraction of users it was possible, so the results may not be reflecting the real patterns on the whole population. The fact that young users tend to be a minority in our results could be due also to the fact that an accurate prediction (i.e. a prediction with accuracy higher than 90%) is possibly harder for young users, and so young users are more likely to remain unclassified.

The gender composition points out a higher presence of men in the conversations; interestingly, this result holds, although to a lesser extent, also for the case of feminism, a topic on which one could expect women to be more active. Italy and Spain present a different scenario: the presence of women is comparable to the presence of men in the case of climate change (only slightly lower), and higher in the case of feminism. The presence of women is comparable to the presence of men also in the conversations on Feminism in the networks of UK and Switzerland.

Looking at the tables, we see in some cases negative values for homophily. This indicates a preference for interaction with users having different characteristics; however, we can notice that when values are negative they are usually quite small, indicating a mostly neutral preference, with just a slight preference for interacting with the other gender/age class. Instead, in several cases we observe higher values of homophily with positive sign.

As a general trend, we observe higher homophily for women, in line with previous literature from other social networks (Laniado et al, 2016). This is true in particular in the French, German, Italian, UK and Swedish networks, where homophily is generally quite high for women, and lower for men. Interestingly, in the countries were the presence of women is higher, and where they get to be a majority in the conversations on Feminism (Italy and Spain) the homophily among men and women is comparable, and close to a neutral preference. In the Spanish conversation networks on Feminism, men are a minority and tend to have a higher homophily, where women exhibit neutral preference for interaction with men or women.

Country	network	% women	% men	homophily women	homophily men
France	retweet	18.4%	36.7%	-0.0422	-0.1518
	mention	15.4%	27.6%	0.0311	-0.0250
Germany	retweet	19.7%	40.1%	-0.0507	-0.1384
	mention	13.6%	40.1%	0.1792	-0.1837
Greece	retweet	13.2%	23.1%	0.0449	0.4810
	mention	6.7%	14.1%	0.1846	0.2044
Italy	retweet	27.9%	30.7%	0.0338	-0.0028
	mention	24.0%	30.8%	0.2231	-0.1127
Poland	retweet	12.8%	17.6%	0.1508	0.3862
	mention	4.3%	8.6%	0.1451	0.2077
Spain	retweet	20.4%	25.5%	-0.1446	-0.1779
	mention	3.1%	4.7%	0.0195	0.0659
Sweden	retweet	28.2%	39.2%	0.4317	-0.0980
	mention	21.3%	45.4%	0.3332	-0.1557
Switzerland	retweet	24.3%	37.0%	-0.0146	-0.0609
	mention	21.9%	38.4%	0.0680	-0.0315
United Kingdom	retweet	27.7%	37.3%	0.0800	-0.0092
	mention	25.7%	38.6%	0.1201	-0.0717

Table 6. **Homophily by gender - Climate change networks.** The table reports for each network the percentage of users identified as below and above the age of 30, and the homophily within each of the two user groups. Positive values of homophily indicate a tendency of users in that group to interact preferentially with each other; negative values a tendency to interact preferentially with users from the other group.

Country	network	% women	% men	homophily women	homophily men
France	retweet	18.9%	36.4%	-0.0726	-0.1529
	mention	15.9%	28.1%	0.1212	-0.0789
Germany	retweet	18.1%	32.2%	0.1036	-0.0534
	mention	17.5%	30.5%	0.0476	-0.0893
Greece	retweet	17.4%	24.7%	0.2860	0.0684
	mention	7.2%	21.7%	0.3297	0.1845
Italy	retweet	30.6%	26.9%	-0.0237	-0.1258
	mention	27.8%	23.1%	-0.0318	-0.0331
Poland	retweet	20.4%	29.1%	0.4573	0.2859
	mention	4.6%	11.8%	0.0045	0.3469
Spain	retweet	24.3%	21.2%	0.0222	0.2158
	mention	15.1%	12.4%	-0.0335	0.2027
Sweden	retweet	20.4%	50.2%	0.1713	0.0782
	mention	21.7%	50.4%	0.1242	0.1738
Switzerland	retweet	30.0%	32.5%	-0.0559	0.0138
	mention	26.5%	31.3%	0.0072	-0.0866
United Kingdom	retweet	30.4%	33.4%	0.2362	-0.0294
	mention	30.1%	33.7%	0.0516	-0.0139

Table 7. **Homophily by gender - Feminism networks.** The table reports for each network the percentage of users identified as below and above the age of 30, and the homophily within each of the two user groups. Positive values of homophily indicate a tendency of users in that group to interact preferentially with each other; negative values a tendency to interact preferentially with users from the other group.

Country	network	% below 30	% above 30	homophily below 30	homophily above 30
France	retweet	8.0%	18.2%	-0.0352	-0.0272
	mention	3.4%	14.9%	-0.0333	0.0518
Germany	retweet	5.5%	14.7%	0.0009	0.0510
	mention	2.7%	15.5%	0.2028	-0.0180
Greece	retweet	1.6%	5.0%	0.1188	0.0745
	mention	0.4%	2.3%	-0.0040	0.0905
Italy	retweet	3.0%	6.7%	0.0754	0.0998
	mention	1.5%	6.7%	0.1537	0.0412
Poland	retweet	1.3%	4.1%	0.3239	0.2878
	mention	0.2%	1.0%	-0.0022	0.0419
Spain	retweet	4.9%	17.3%	-0.0403	-0.0726
	mention	0.3%	4.3%	-0.0032	0.0826
Sweden	retweet	5.8%	15.0%	0.6547	0.0728
	mention	1.0%	10.8%	0.4230	0.0468
Switzerland	retweet	2.6%	12.1%	0.0716	0.0240
	mention	1.5%	10.6%	0.1531	-0.0040
United Kingdom	retweet	1.5%	4.7%	0.0748	0.1023
	mention	0.6%	4.8%	0.2017	0.0564

Table 8. **Homophily by age - Climate change networks.** The table reports for each network the percentage of users identified as below and above the age of 30, and the homophily within each of the two user groups. Positive values of homophily indicate a tendency of users in that group to interact preferentially with each other; negative values a tendency to interact preferentially with users from the other group.

Country	network	% below 30	% above 30	homophily below 30	homophily above 30
France	retweet	8.6%	18.0%	-0.0535	-0.0409
	mention	3.1%	17.1%	0.0377	-0.0127
Germany	retweet	4.9%	16.7%	0.0823	0.1711
	mention	4.1%	18.1%	-0.0322	-0.0070
Greece	retweet	5.0%	10.0%	0.2239	0.0673
	mention	0.7%	9.6%	NULL	0.1920
Italy	retweet	9.6%	17.4%	-0.0744	0.1076
	mention	5.5%	18.3%	0.0834	0.0229
Poland	retweet	7.5%	12.4%	-0.0603	0.3266
	mention	1.2%	3.9%	0.2384	0.1620
Spain	retweet	11.3%	13.1%	-0.0989	0.3226
	mention	2.7%	11.6%	0.0890	0.2442
Sweden	retweet	3.7%	25.8%	-0.0367	-0.0424
	mention	2.6%	24.7%	-0.0246	-0.0263
Switzerland	retweet	5.0%	21.6%	0.0066	0.0696
	mention	3.4%	21.6%	-0.0327	-0.0178
United Kingdom	retweet	6.2%	16.5%	0.0853	0.0926
	mention	3.8%	19.6%	0.0409	0.0619

Table 9. **Homophily by age - Feminism networks.** The table reports for each network the percentage of users identified as below and above the age of 30, and the homophily within each of the two user groups. Positive values of homophily indicate a tendency of users in that group to interact preferentially with each other; negative values a tendency to interact preferentially with users from the other group.

3.3. Inequality analysis results by gender and age range

In the following, we present the distributions of two variables, number of tweets and in-degree, comparing different classes of users, namely comparing users by gender and by age class.

The distribution of the number of tweets by user gives an indication of the involvement of users in the debate. This distribution, according to a very common rule in the internet, usually follows a heavy tailed pattern, with most of the users having a very low activity, i.e. they just posted one or few tweets, and a few users having higher activity levels, up to hundreds or even thousands of tweets posted by a single user in our dataset. Therefore we observe in the log-log graphs (with logarithmic scale on both axes) the typical shape that tends to follow a power law, with high values on the left of the graph (many users with low activity) and decreasing values towards the right of the graph, indicating less and less users with high levels of activity. This decreasing line may be more or less skewed, indicating higher or lower inequality.

In the graphs showing the in-degree distributions in the networks of retweets and mentions we see a similar pattern: most users (on the left side of the graphs) receive attention in the form of a retweet or a mention from just one or few other users, while a few users (towards the right of the graphs) receive attention from many other users, and have higher levels of centrality. In this case we find less users than in the activity graphs, because we show only users receiving at least one mention/retweet from some other user in the dataset: we exclude users that appear in our dataset because they post some tweet, but do not receive attention from other users.

In both cases, when representing activity and centrality in terms of attention received by other users, we highlight inequalities according to demographic characteristics of the users by representing different classes of users in different colors and symbols. This is the convention we follow consistently in all the graphs:

- When comparing users by age class:
    - green circles represent users identified as below 30 years old
    - purple triangles represent users identified as above 30 years old
    - gray crosses represent users whose age could not be inferred

- When comparing users by gender:
    - red circles represent users identified as women
    - blue crosses represent users identified as men
    - gray triangles represent users whose gender could not be inferred.

For deepening into the inequalities among young users, for the in-degree distributions we further show the distributions by gender among users below the age of 30; i.e., we plot the same graph, restricted to users that were reliably identified as younger than 30. Due to the difficulty of reliably identifying the age class for most users, only a minority of users are sown in this case (not only

older users are filtered out, but also users for whom it was not possible to identify the age class with due accuracy). When the amount of users identified as younger than 30 was too low (only a few individuals) we omitted these filtered graphs.

In the following, for each country and for each topic, first we show the distribution of activity by age class and gender, which gives an indication of how involved in the debate are different classes of users; then we look at the attention that users with different demographic characteristics receive in the networks of retweets and mentions.

3.3.1. France

3.3.1.1. ClimateStrike

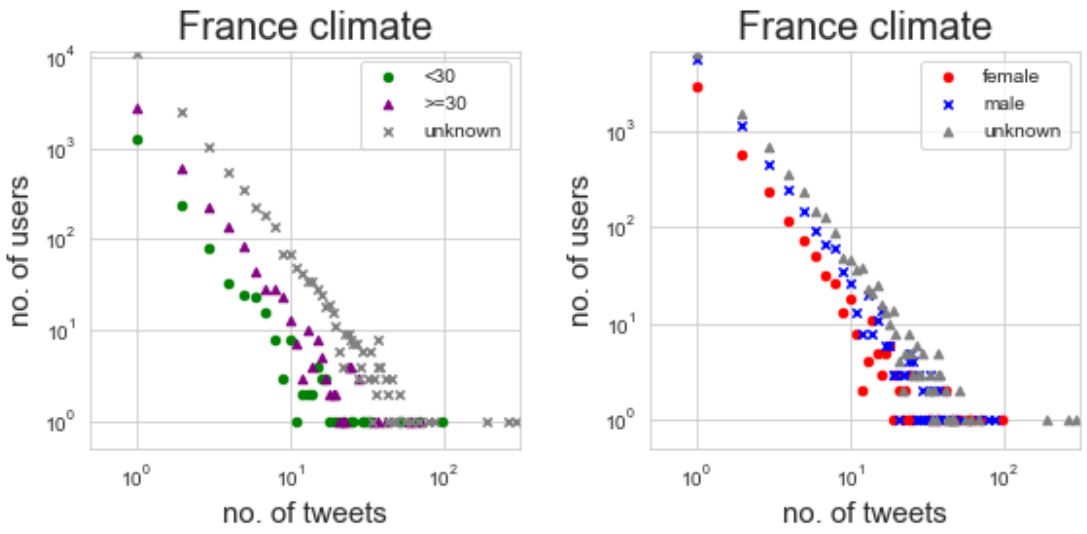

Figure 10. Distributions of users by number of tweets in France about climate change considering age range (left) and gender (right)

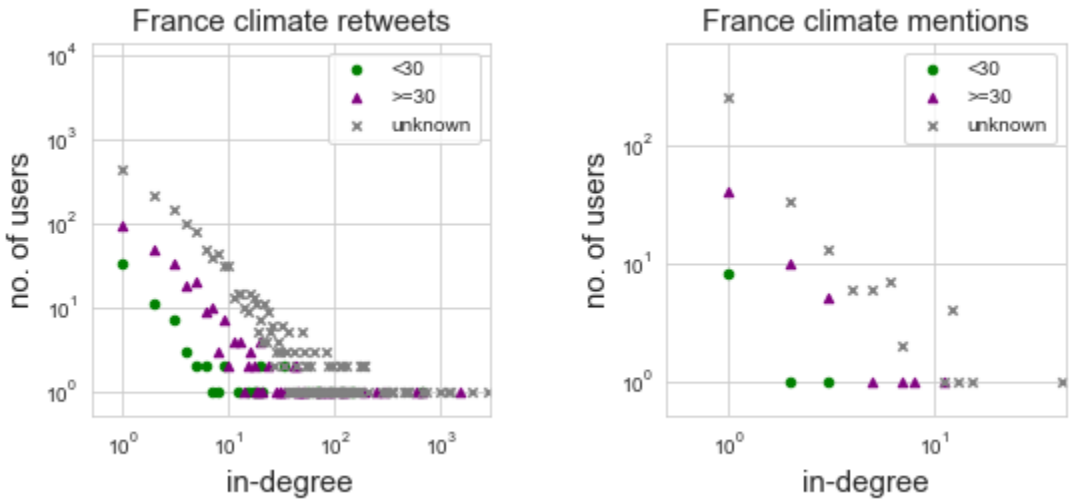

Figure 11. Indegree distribution by age in the French networks for Climate change based on retweets (left) and mentions (right).

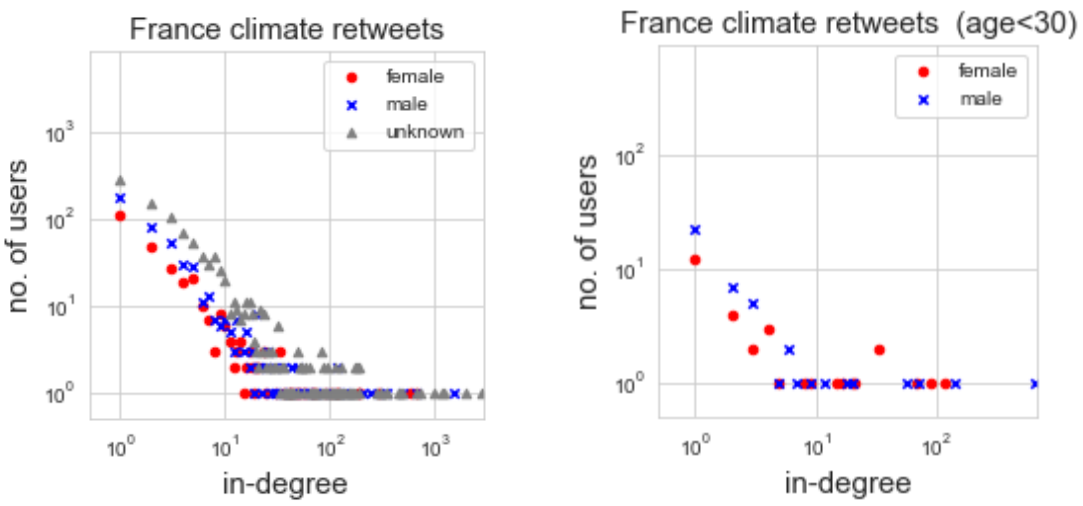

Figure 12. Indegree distribution by gender in the French networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

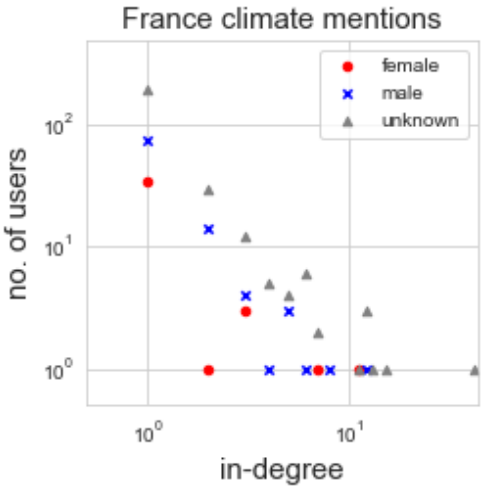

Figure 13. Indegree distribution by gender in the French networks for Climate change based on retweets.

3.3.1.2. Feminism

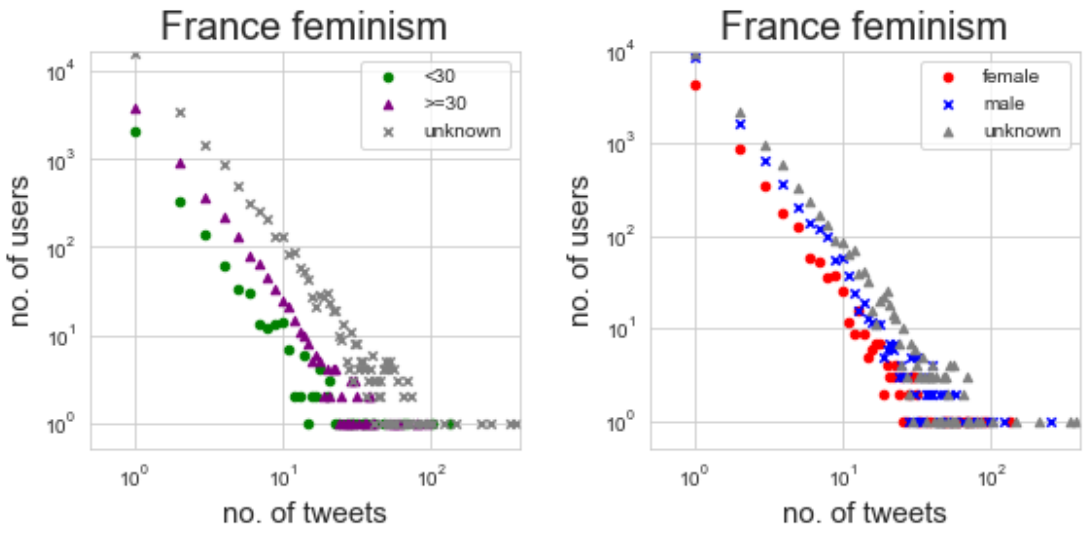

Figure 14. Distributions of users by number of tweets in France about feminism considering age range (left) and gender (right)

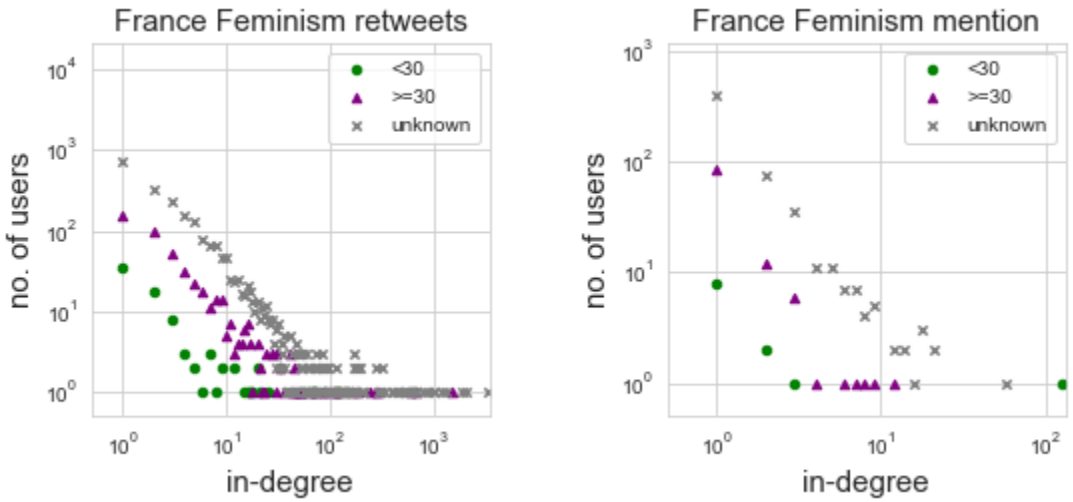

Figure 15. Indegree distribution by age in the French networks for Feminism based on retweets (left) and mentions (right).

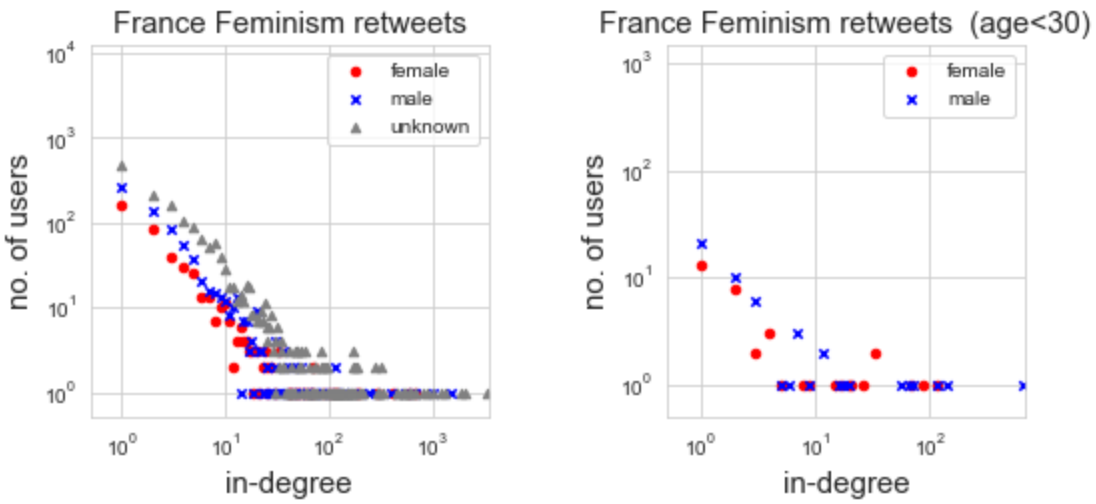

Figure 16. Indegree distribution by gender in the French networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

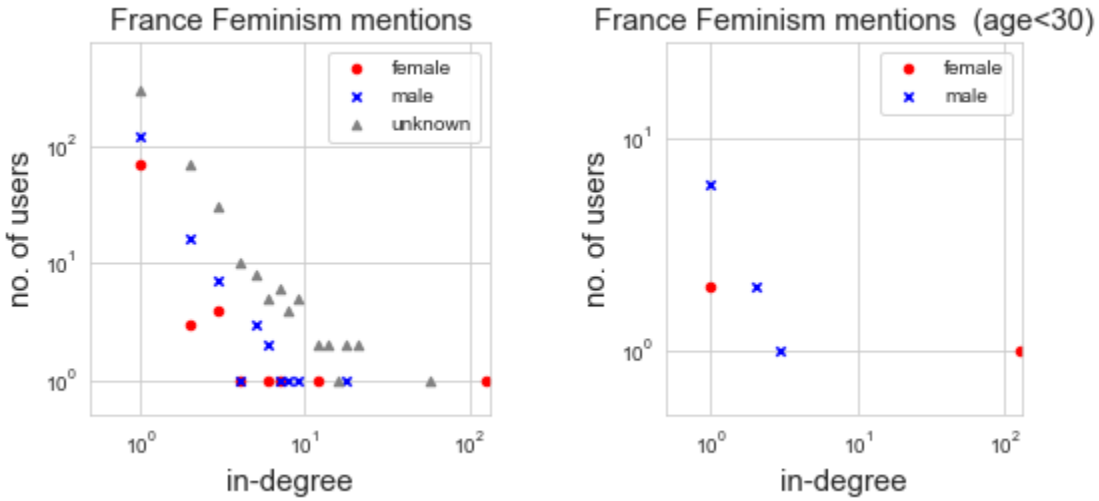

Figure 17. Indegree distribution by gender in the French networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.2. Germany

3.3.2.1. ClimateStrike

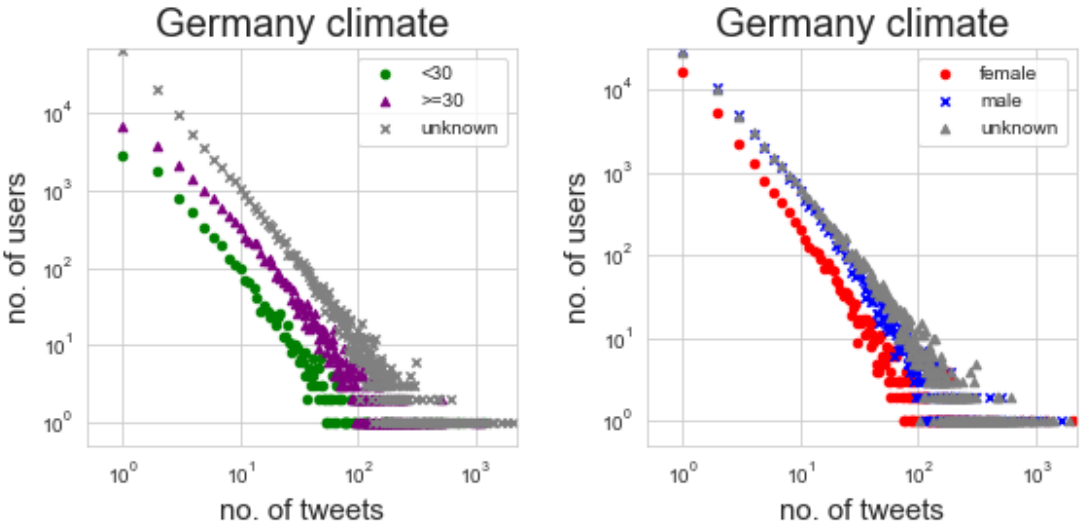

Figure 18. Distributions of users by number of tweets in Germany about climate change considering age range (left) and gender (right)

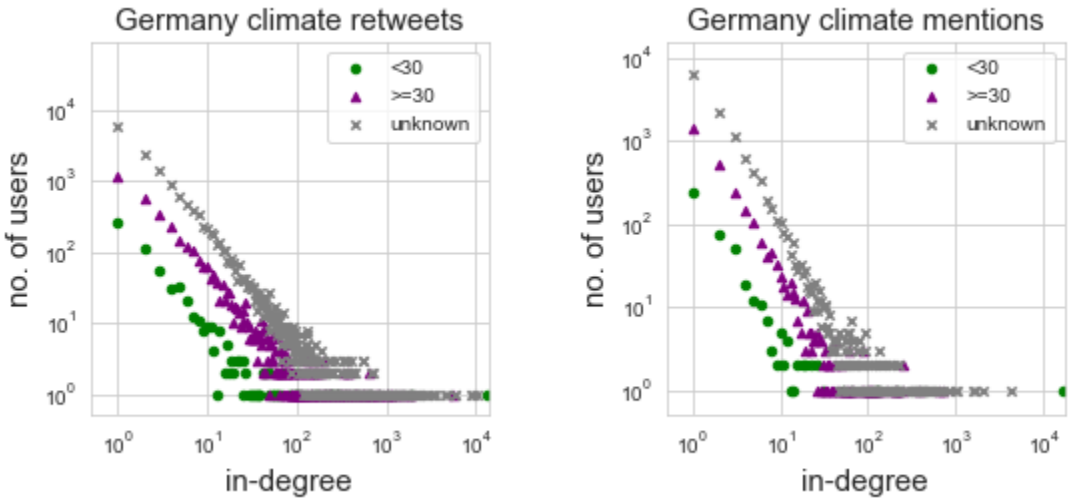

Figure 19. Indegree distribution by age in the German networks for Climate change based on retweets (left) and mentions (right).

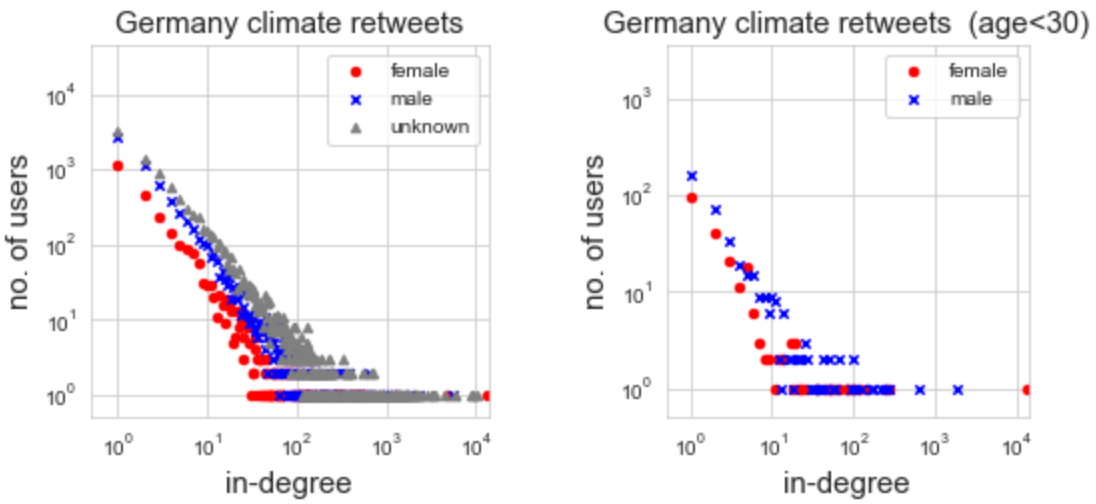

Figure 20. Indegree distribution by gender in the German networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

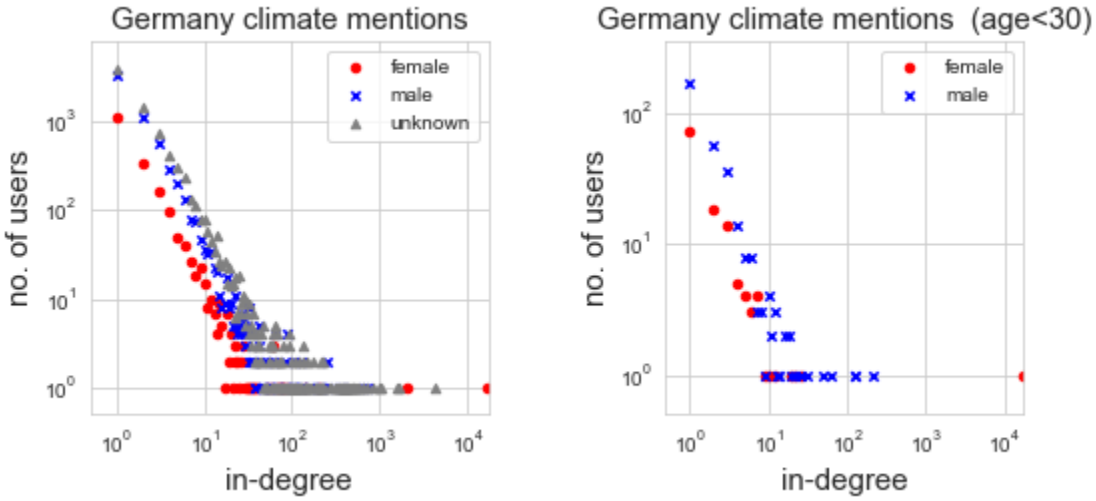

Figure 21. Indegree distribution by gender in the German networks for Climate change based on mentions, for all users (left) and only for users below 30 years old (right).

3.3.2.2. Feminism

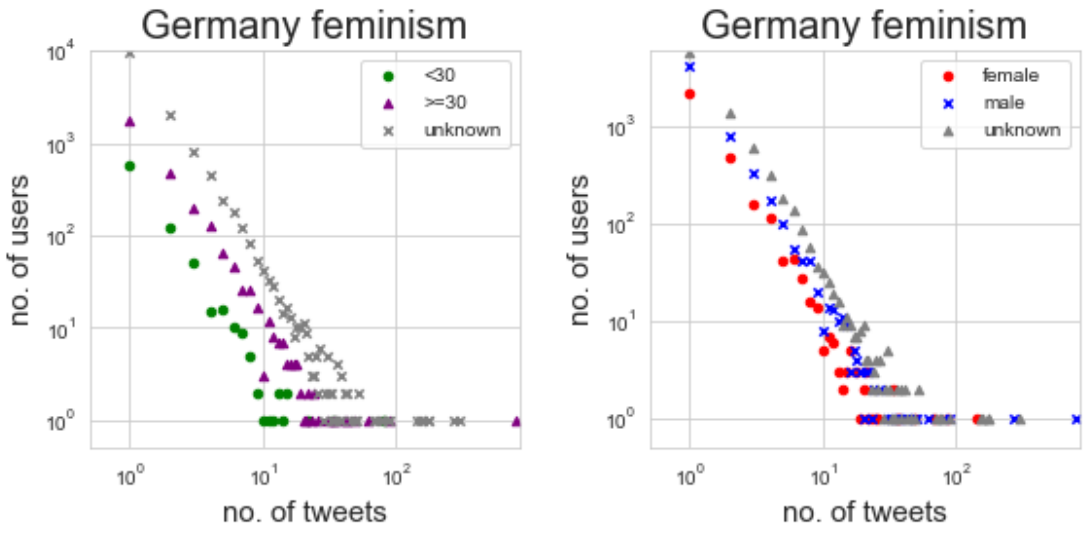

Figure 22. Distributions of users by number of tweets in Germany about feminism considering age range (left) and gender (right)

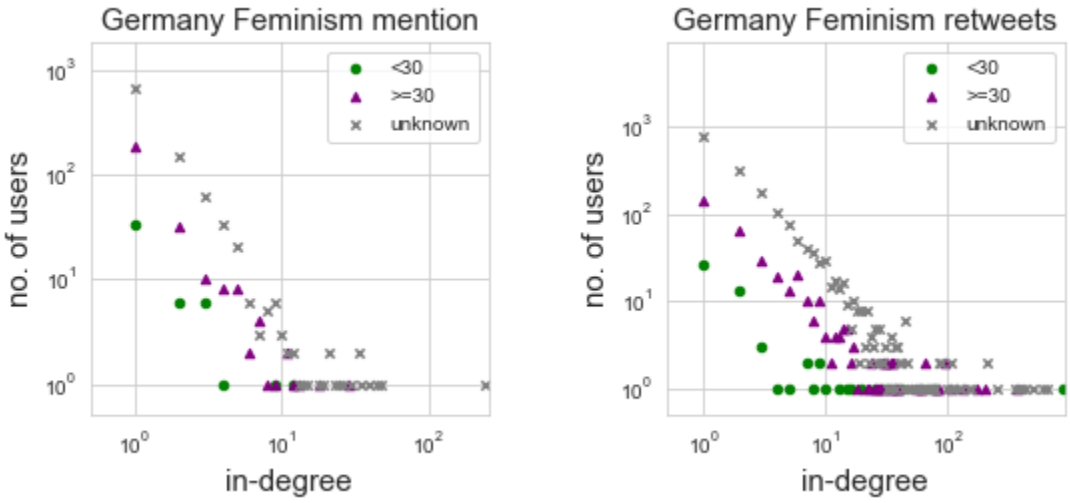

Figure 23. Indegree distribution by age in the German networks for Feminism based on retweets (left) and mentions (right).

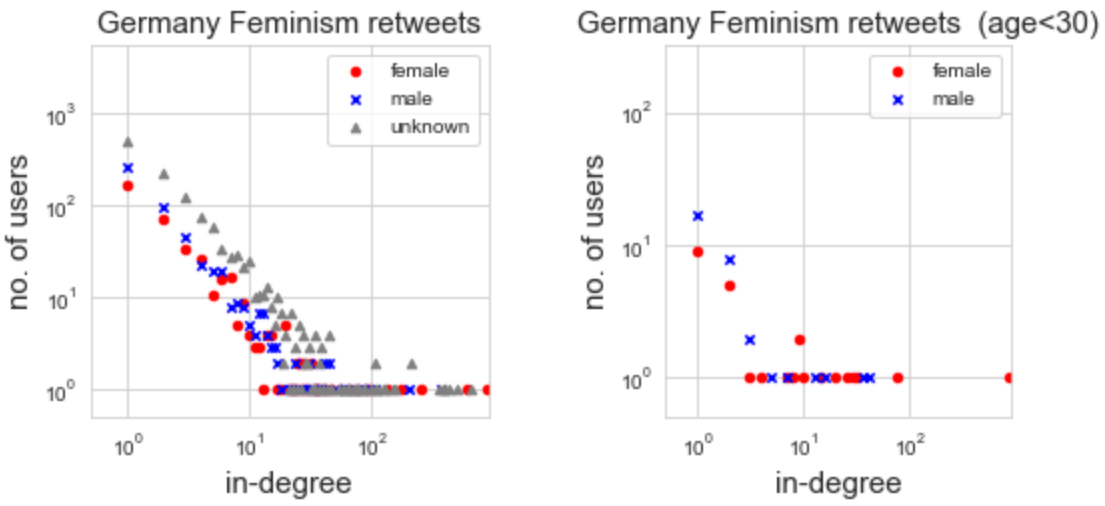

Figure 24. Indegree distribution by gender in the French networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

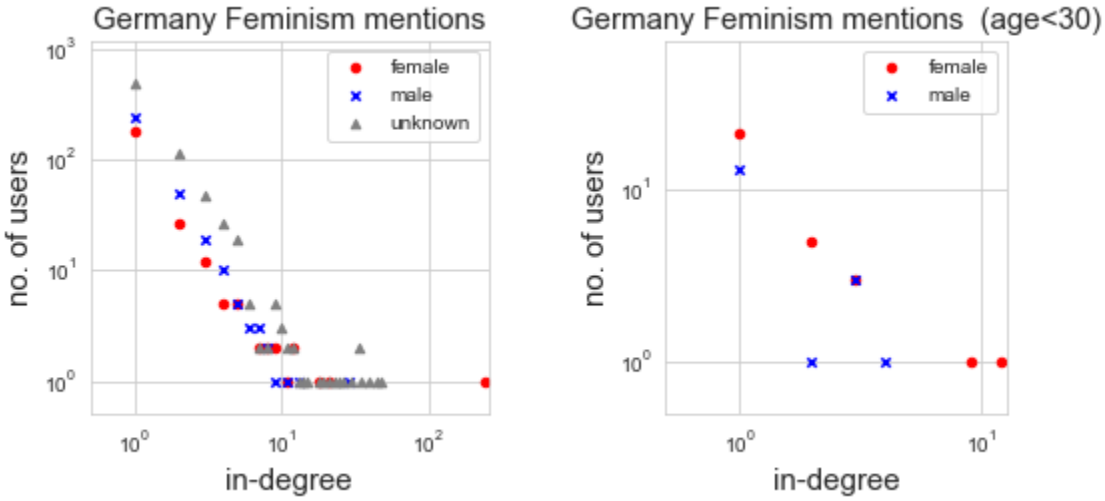

Figure 25. Indegree distribution by gender in the French networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.3. Greece

3.3.3.1. ClimateStrike

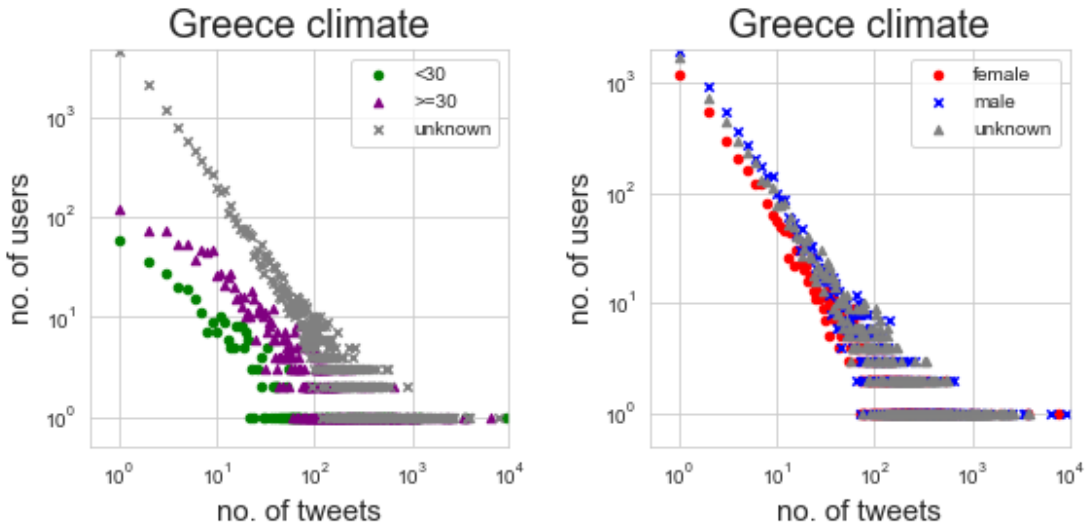

Figure 26. Distributions of users by number of tweets in Greece about climate change considering age range (left) and gender (right)

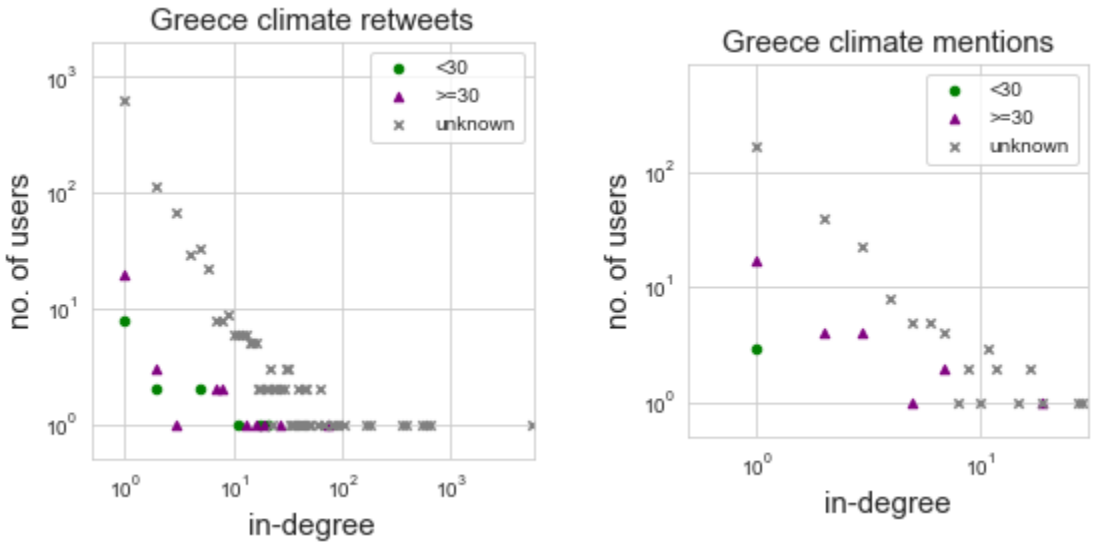

Figure 27. Indegree distribution by age in the Greek networks for Climate change based on retweets (left) and mentions (right).

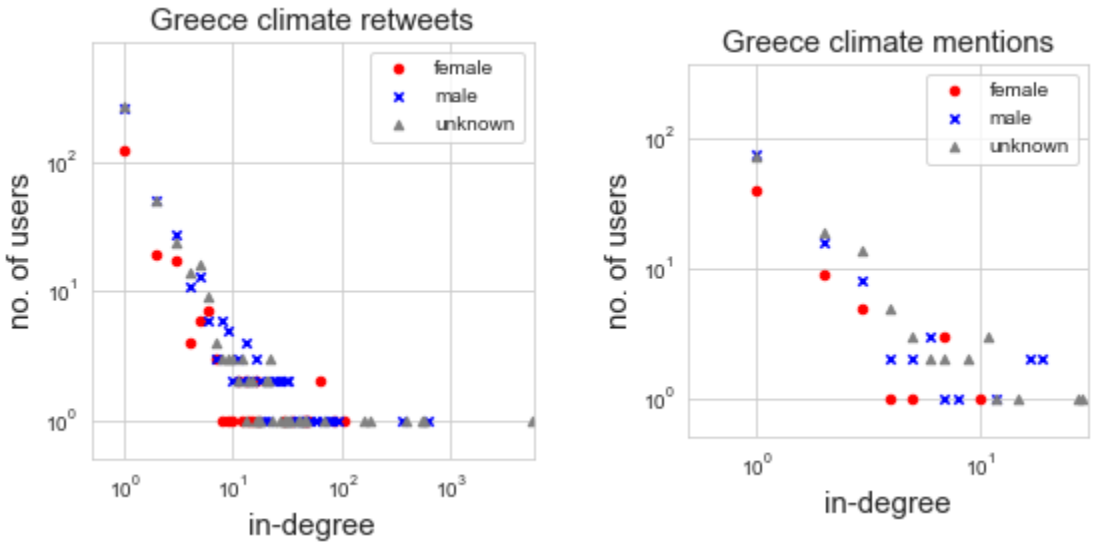

Figure 28. Indegree distribution by gender in the Greek networks for Climate change based on retweets (left) and mentions (right).

3.3.3.2. Feminism

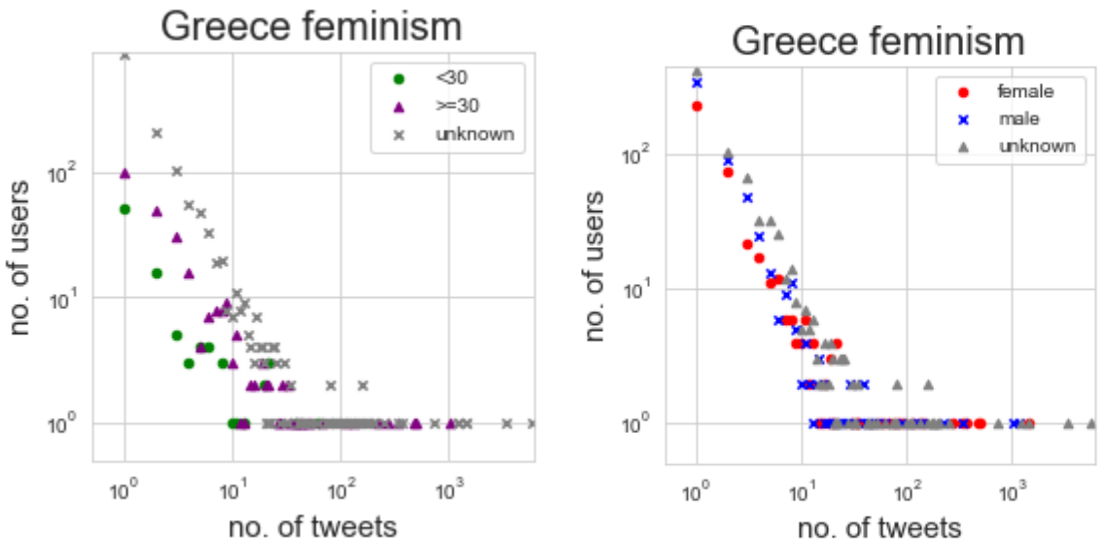

Figure 29. Distributions of users by number of tweets in Greece about feminism considering age range (left) and gender (right)

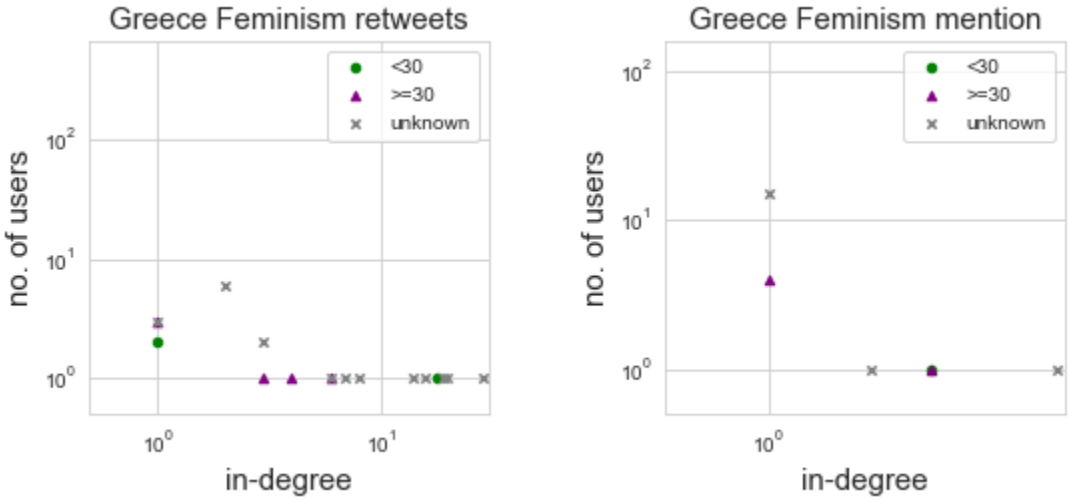

Figure 30. Indegree distribution by age in the Greek networks for Feminism based on retweets (left) and mentions (right).

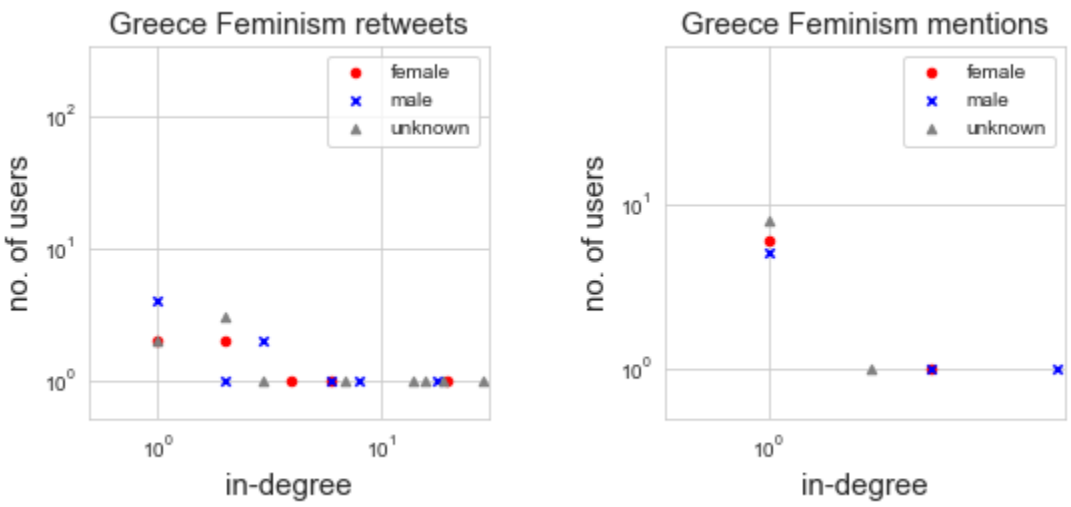

Figure 31. Indegree distribution by gender in the Greek networks for Feminism based on retweets (left) and mentions (right).

3.3.4. Italy

3.3.4.1. ClimateStrike

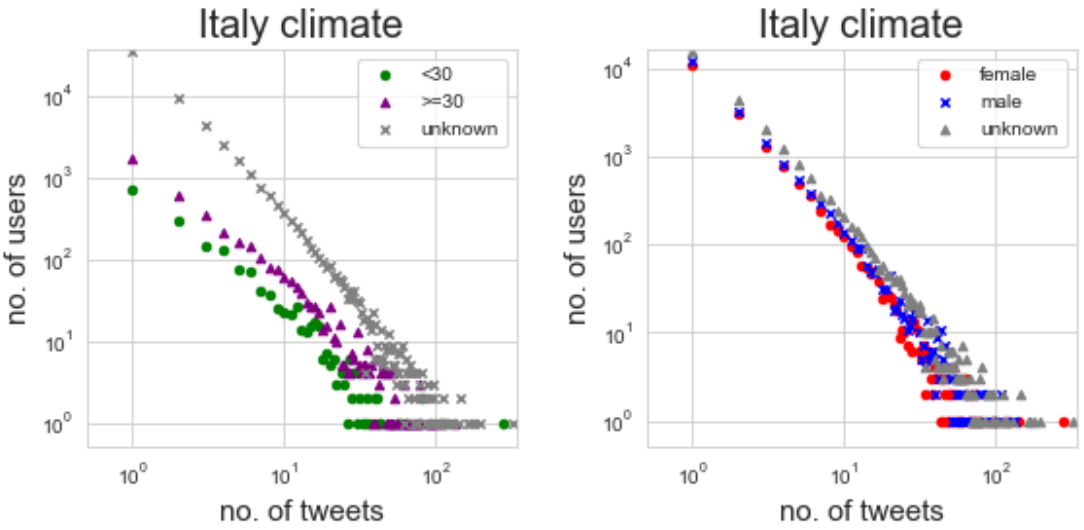

Figure 32. Distributions of users by number of tweets in Italy about climate change considering age range (left) and gender (right)

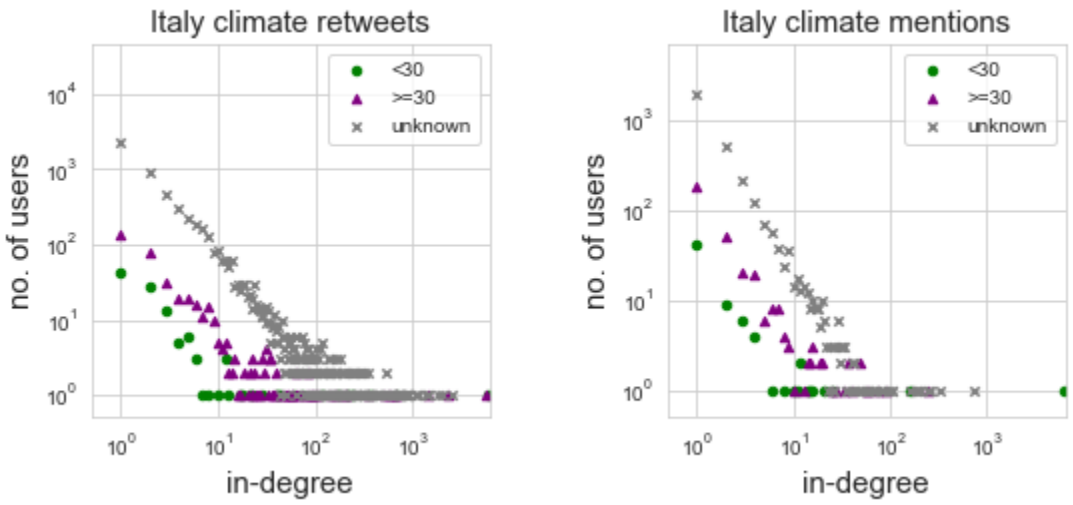

Figure 33. Indegree distribution by age in the Italian networks for Climate change based on retweets (left) and mentions (right).

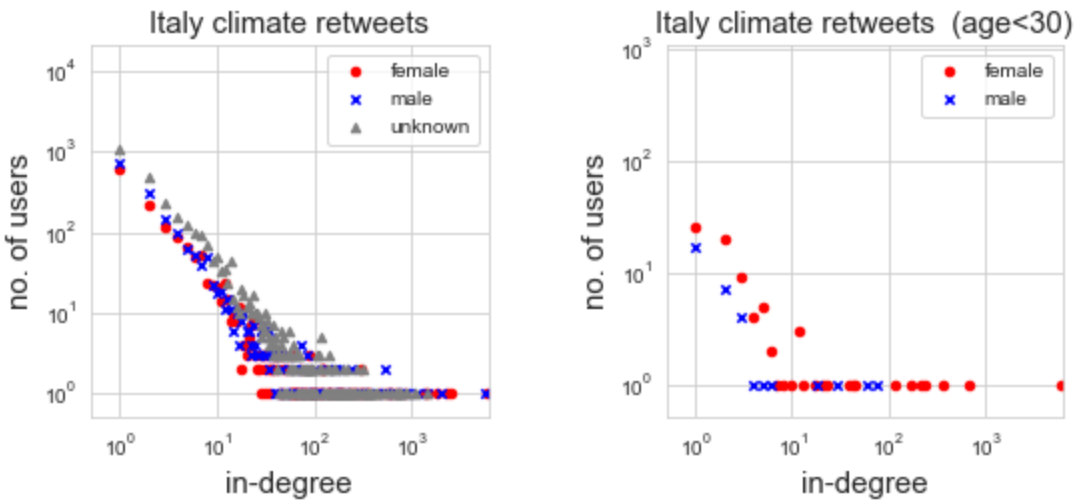

Figure 34. Indegree distribution by gender in the Italian networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

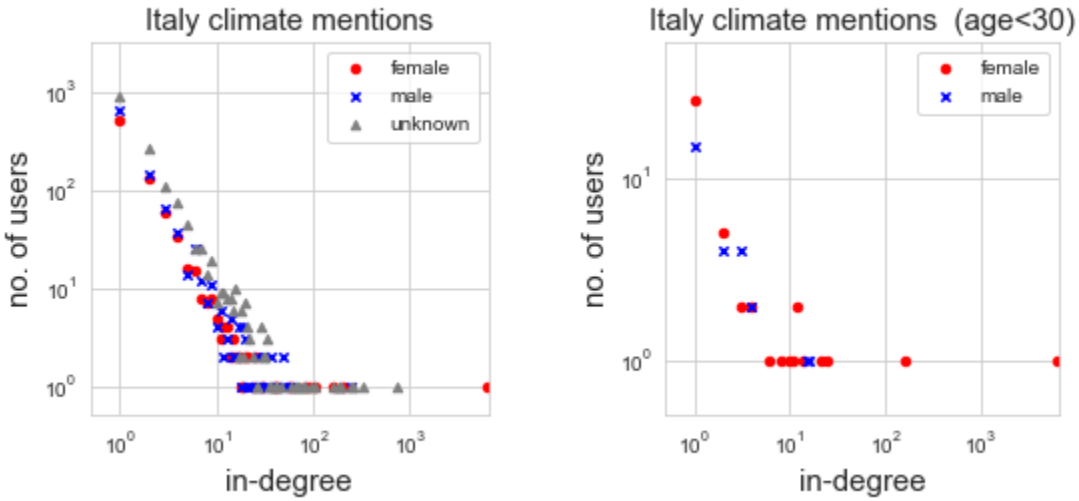

Figure 35. Indegree distribution by gender in the Italian networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.4.2. Feminism

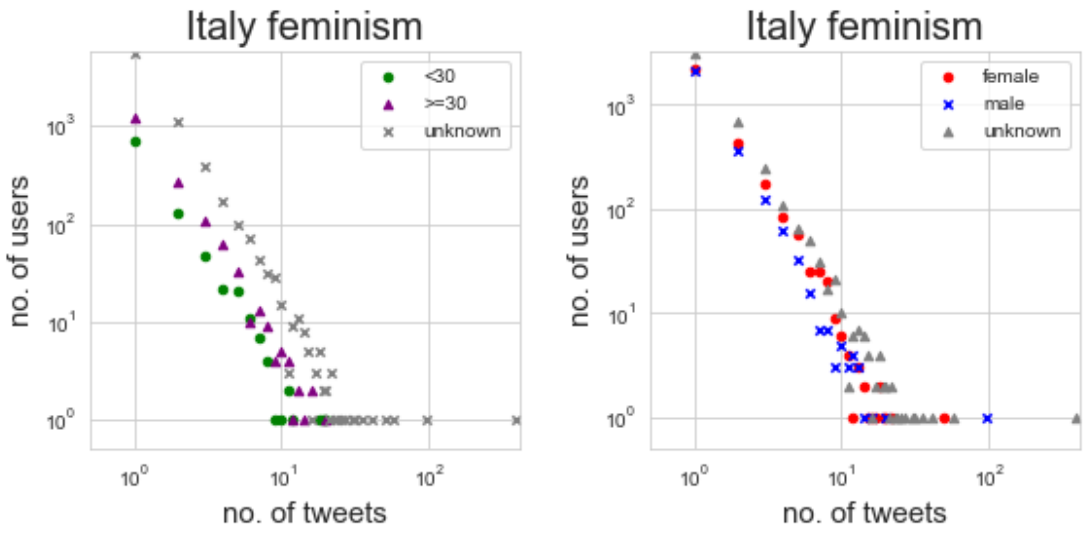

Figure 36. Distributions of users by number of tweets in Italy about feminism considering age range (left) and gender (right)

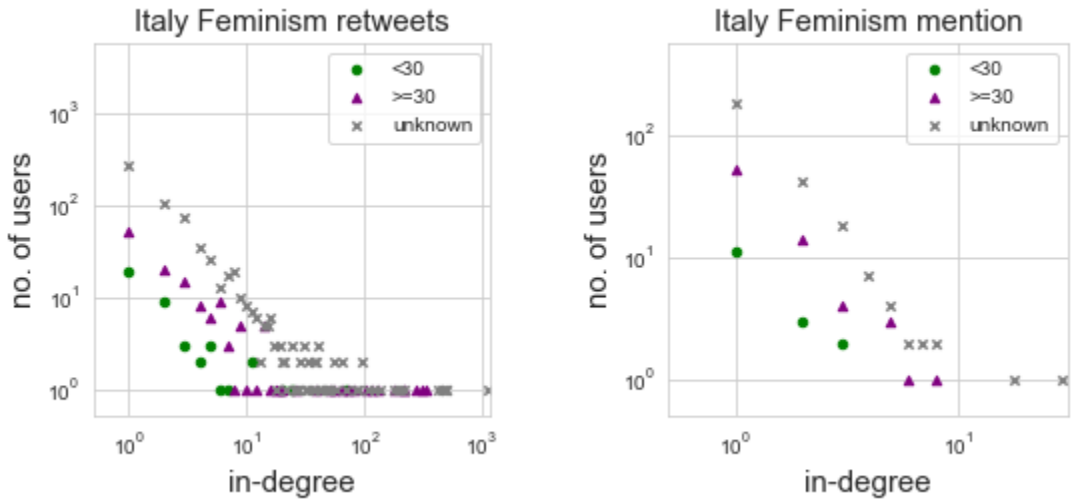

Figure 37. Indegree distribution by age in the Italian networks for Feminism based on retweets (left) and mentions (right).

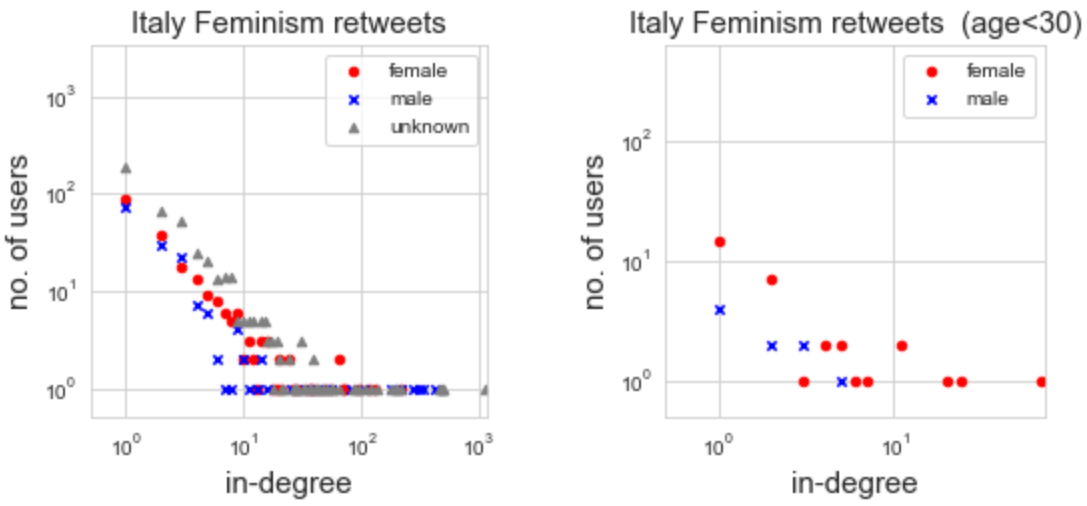

Figure 38. Indegree distribution by gender in the Italian networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

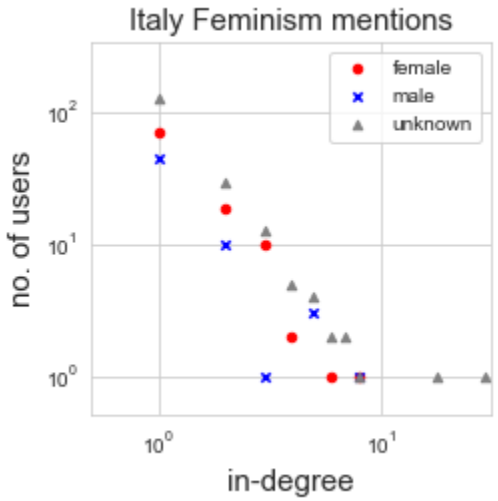

Figure 39. Indegree distribution by gender in the Italian networks for Feminism based on retweets.

3.3.5. Poland

3.3.5.1. ClimateStrike

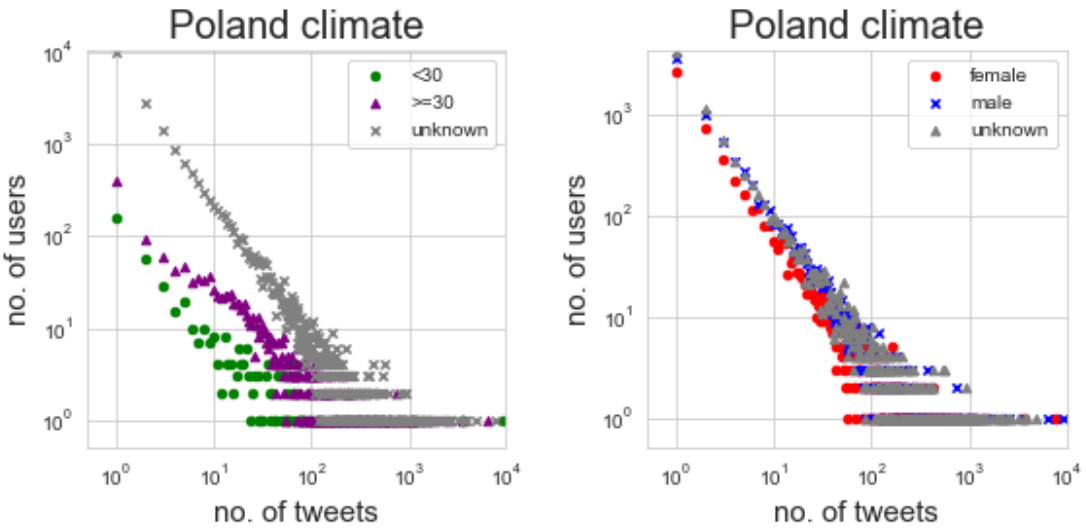

Figure 40. Distributions of users by number of tweets in Poland about climate change considering age range (left) and gender (right)

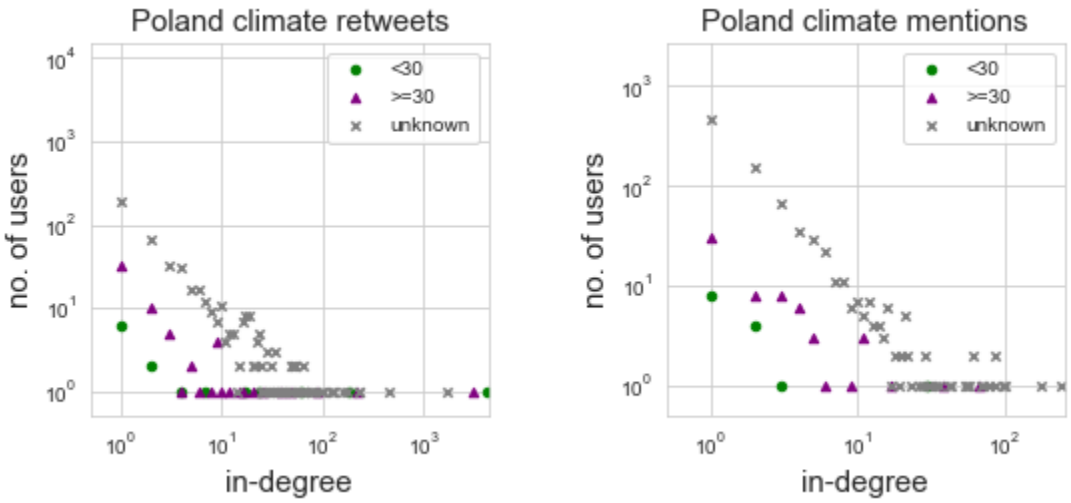

Figure 41. Indegree distribution by age in the Polish networks for Climate change based on retweets (left) and mentions (right).

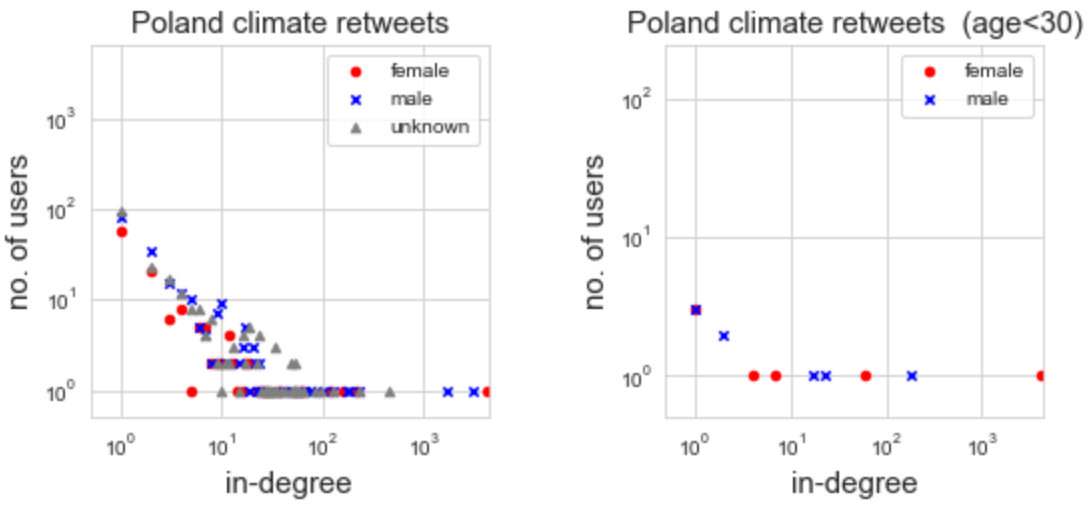

Figure 42. Indegree distribution by gender in the Polish networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

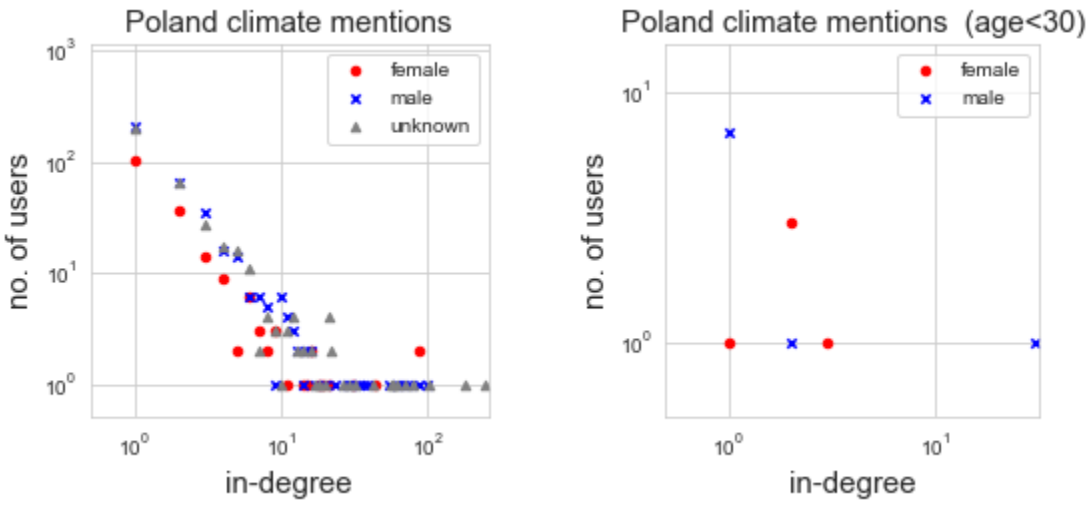

Figure 43. Indegree distribution by gender in the Polish networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.5.2. Feminism

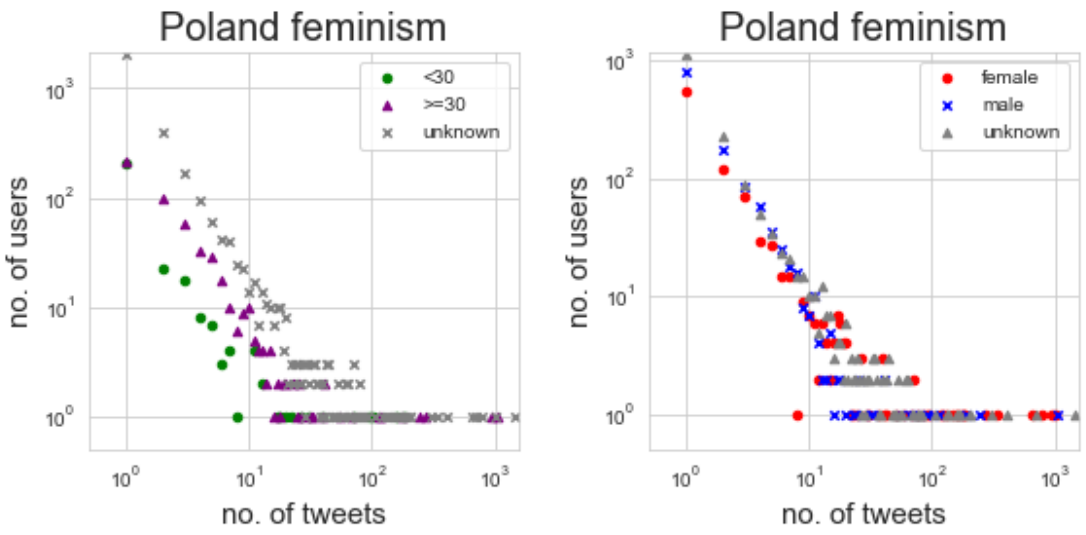

Figure 44. Distributions of users by number of tweets in Poland about feminism considering age range (left) and gender (right)

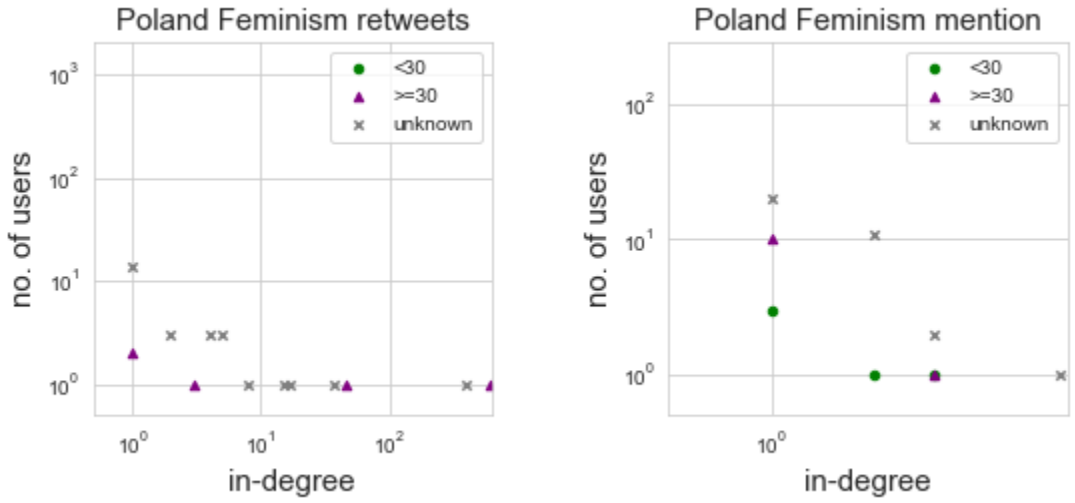

Figure 45. Indegree distribution by age in the Polish networks for Feminism based on retweets (left) and mentions (right).

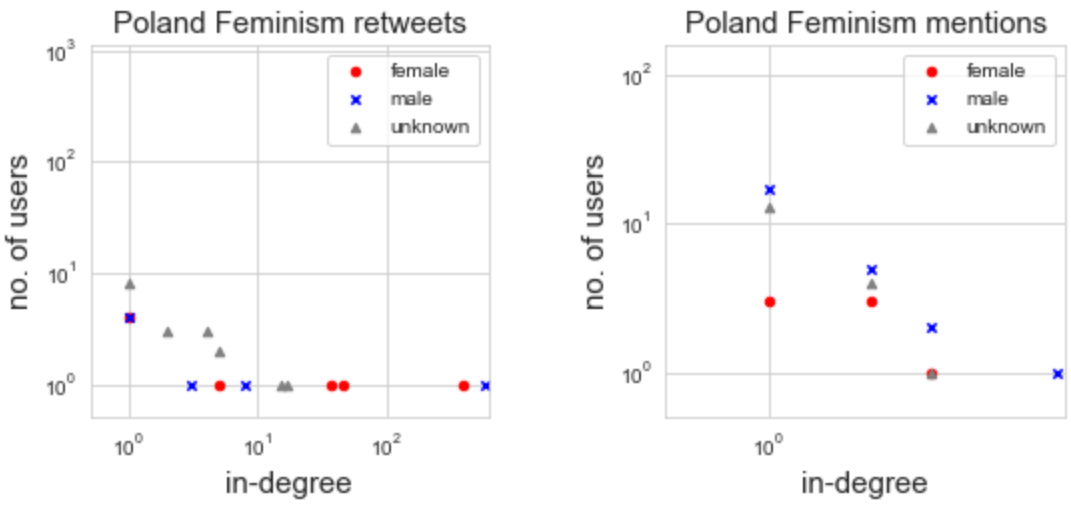

Figure 46. Indegree distribution by gender in the Polish networks for Feminism based on retweets (left) and mentions (right).

3.3.6. Spain

3.3.6.1. ClimateStrike

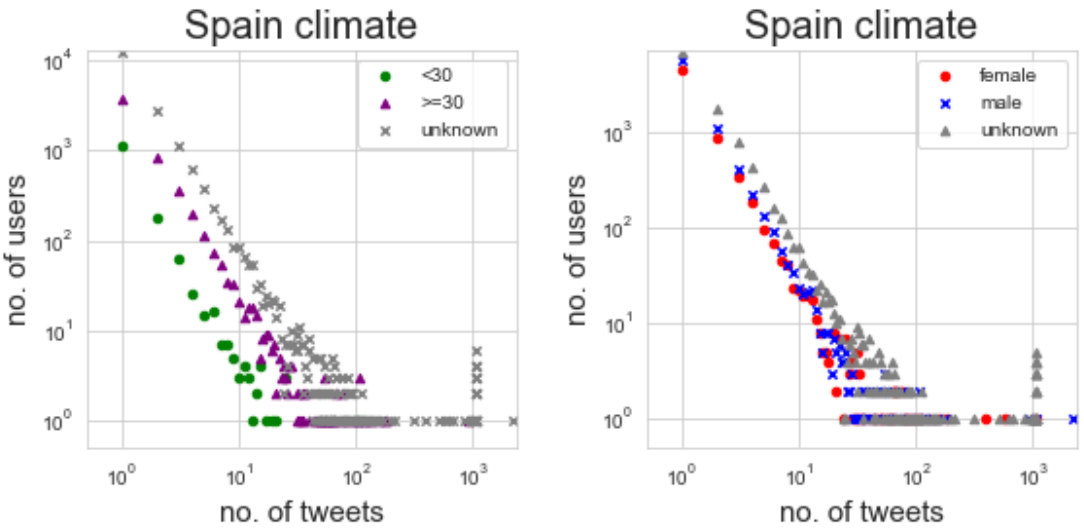

Figure 47. Distributions of users by number of tweets in Spain about climate change considering age range (left) and gender (right)

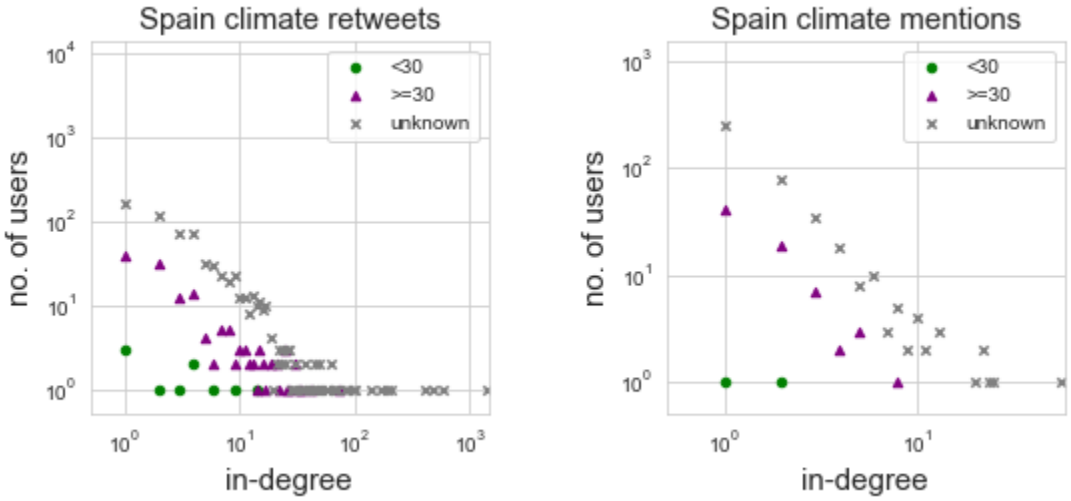

Figure 48. Indegree distribution by age in the Spanish networks for Climate change based on retweets (left) and mentions (right).

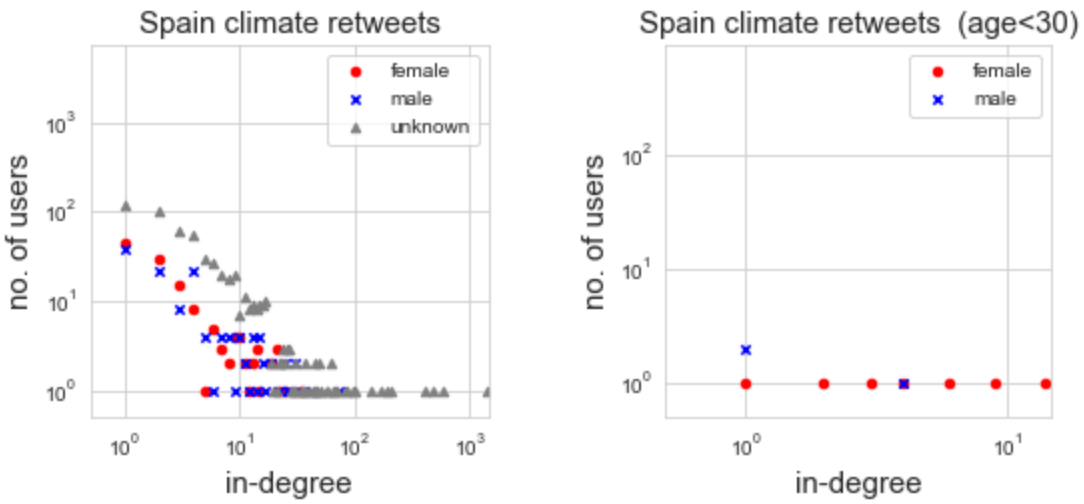

Figure 49. Indegree distribution by gender in the Spanish networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

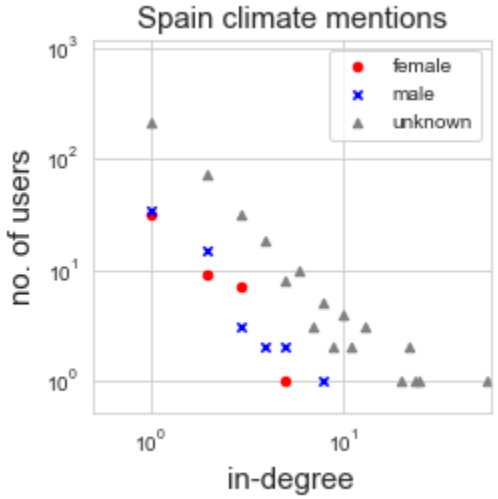

Figure 50. Indegree distribution by gender in the Spanish networks for Climate change based on retweets.

3.3.6.2. Feminism

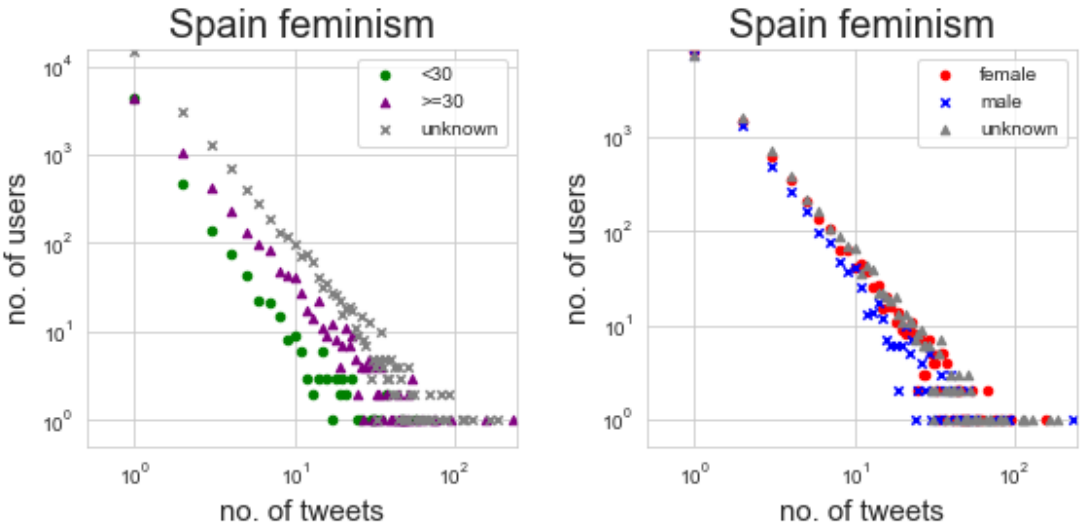

Figure 51. Distributions of users by number of tweets in Spain about feminism considering age range (left) and gender (right)

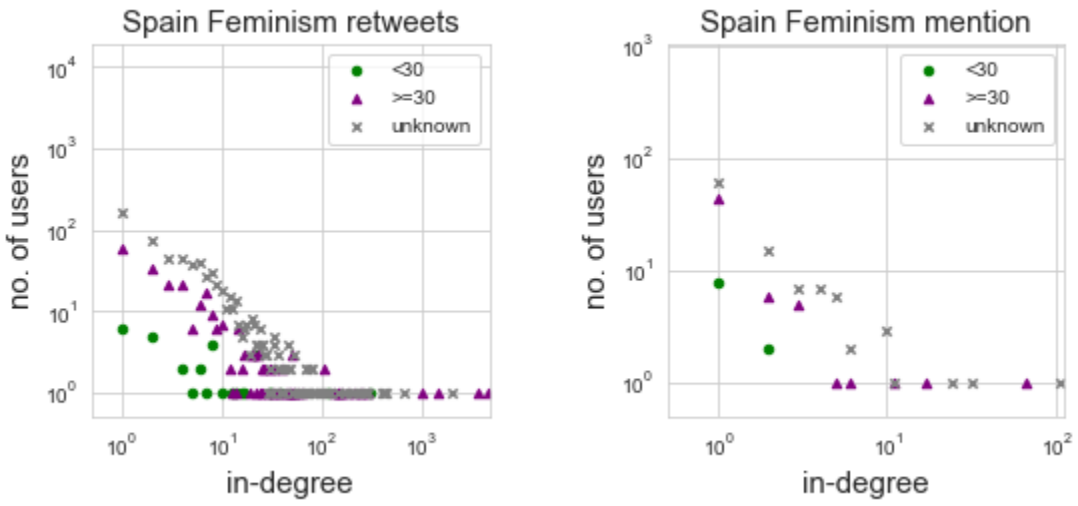

Figure 52. Indegree distribution by age in the Spanish networks for Feminism based on retweets (left) and mentions (right).

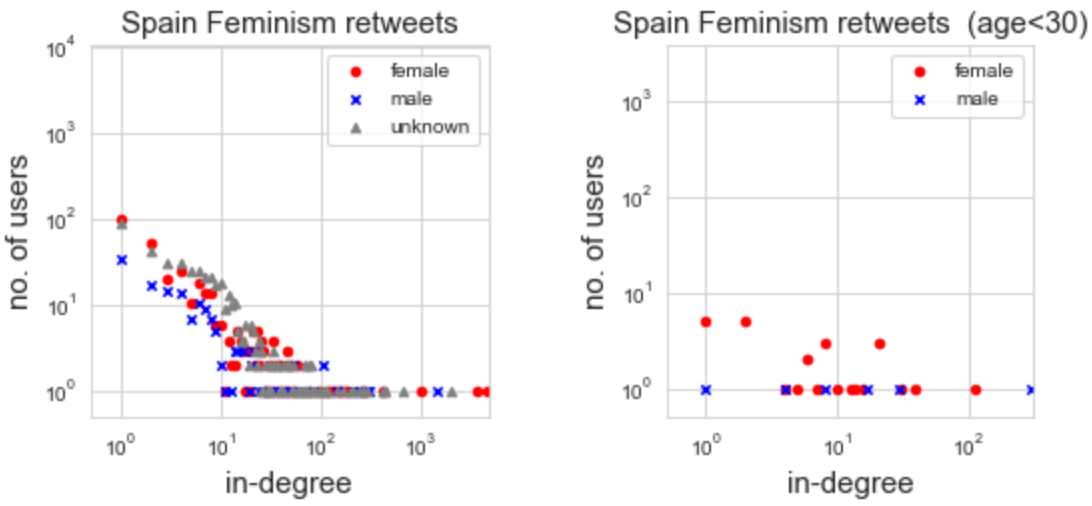

Figure 53. Indegree distribution by gender in the Spanish networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

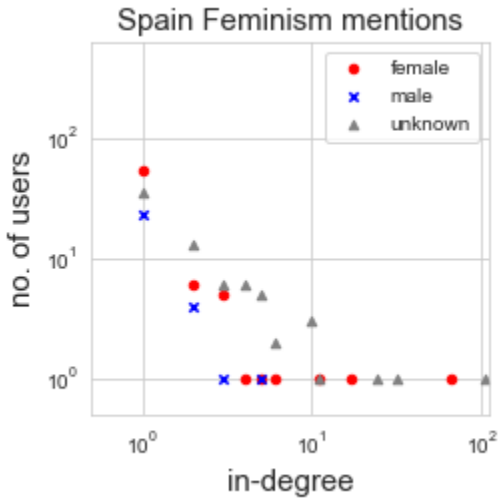

Figure 54. Indegree distribution by gender in the Spanish networks for Feminism based on retweets..

3.3.7. Sweden

3.3.7.1. ClimateStrike

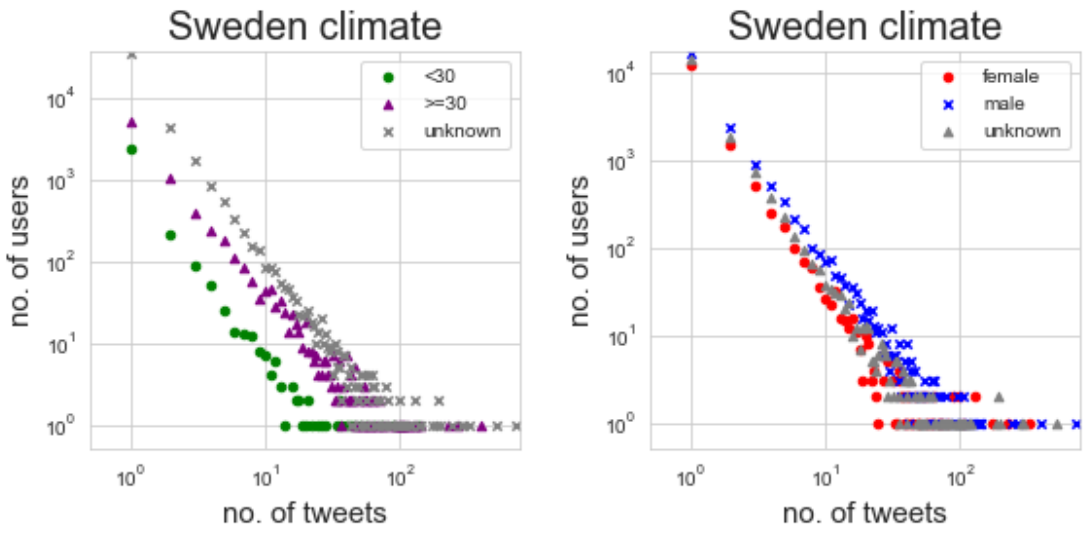

Figure 55. Distributions of users by number of tweets in Sweden about climate change considering age range (left) and gender (right)

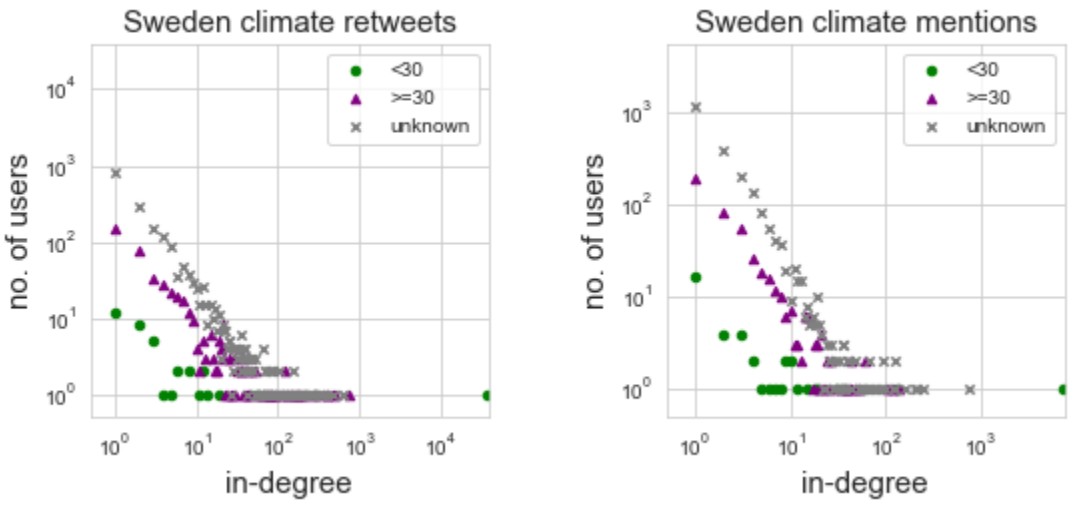

Figure 56. Indegree distribution by age in the Swedish networks for Climate change based on retweets (left) and mentions (right).

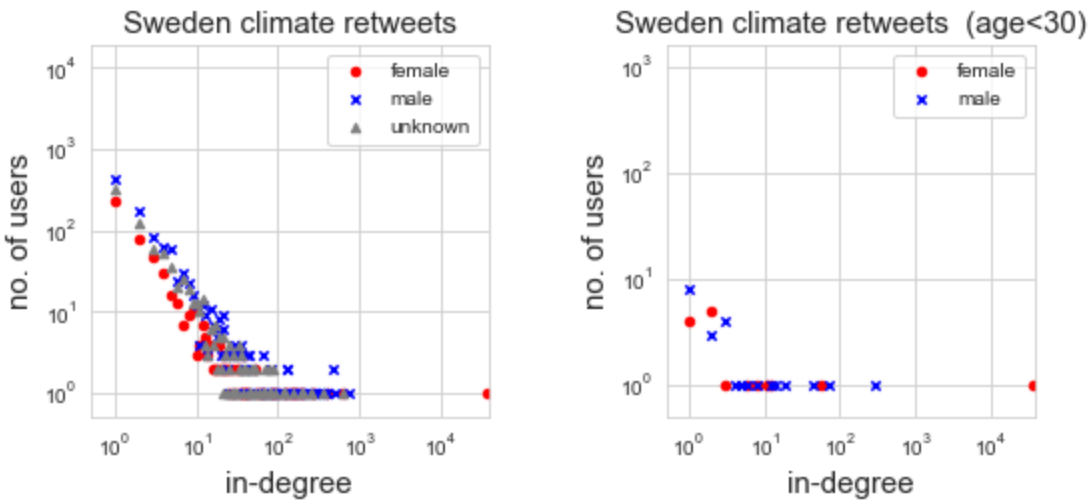

Figure 57. Indegree distribution by gender in the Swedish networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

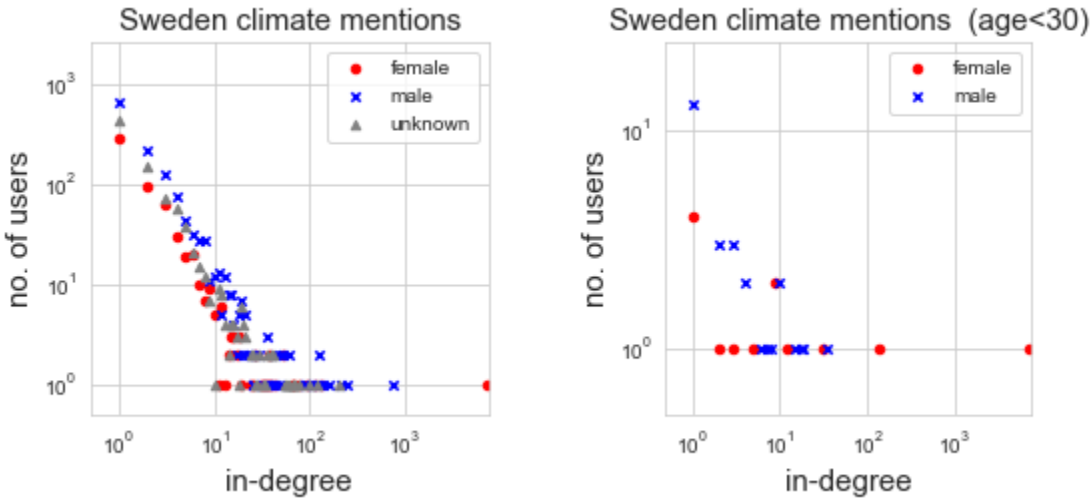

Figure 58. Indegree distribution by gender in the Swedish networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.7.2. Feminism

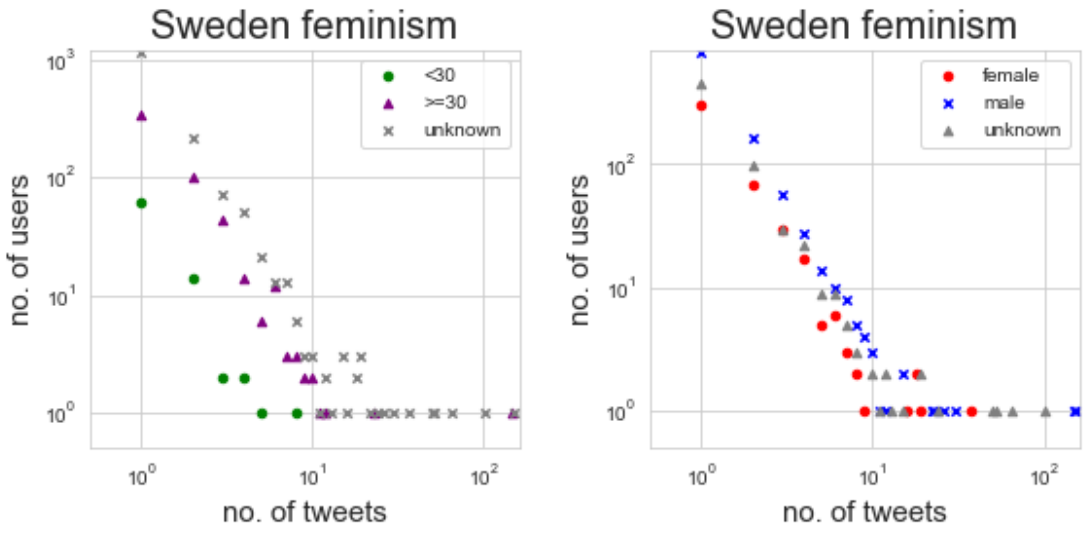

Figure 59. Distributions of users by number of tweets in Sweden about Feminism considering age range (left) and gender (right)

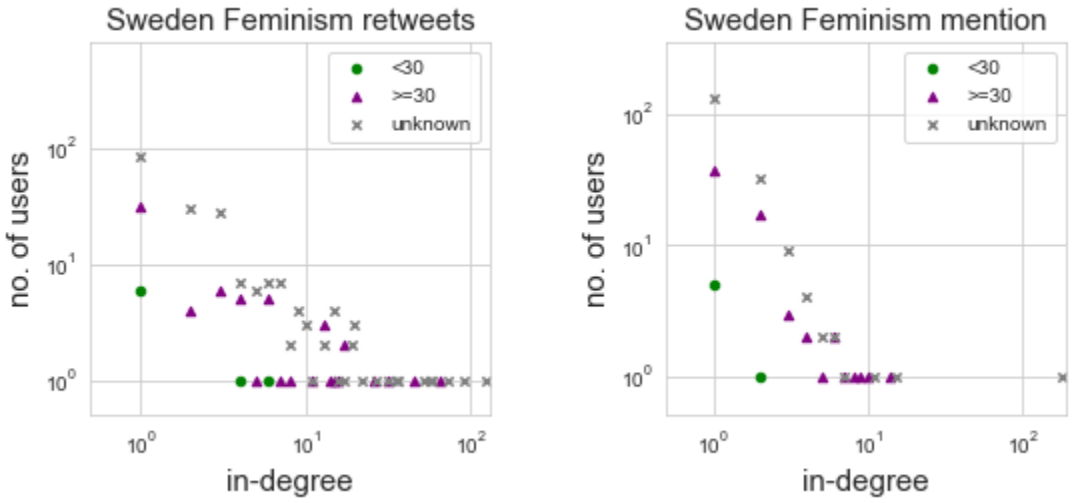

Figure 60. Indegree distribution by age in the Swedish networks for Feminism based on retweets (left) and mentions (right).

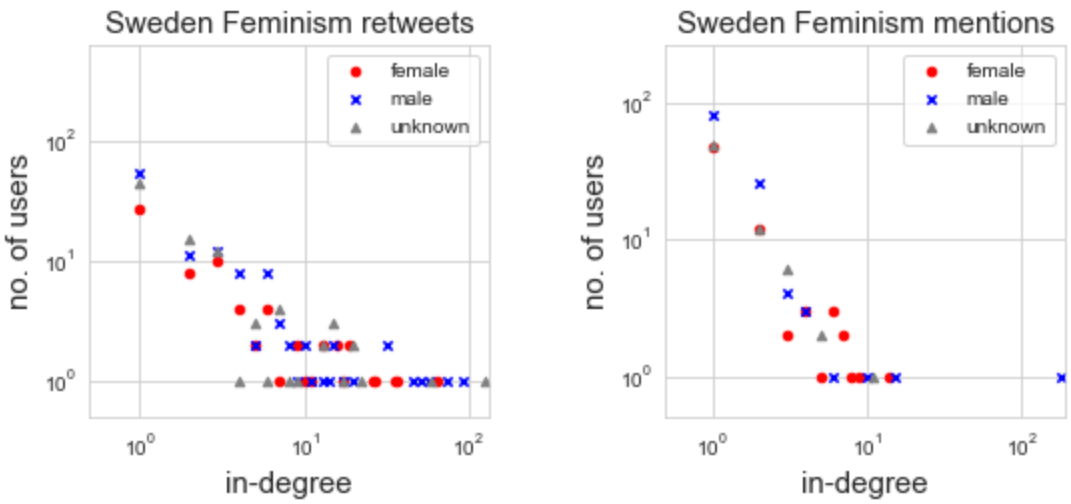

Figure 61. Indegree distribution by gender in the Swedish networks for Feminism based on retweets (left) and mentions (right).

3.3.8. Switzerland

3.3.8.1. ClimateStrike

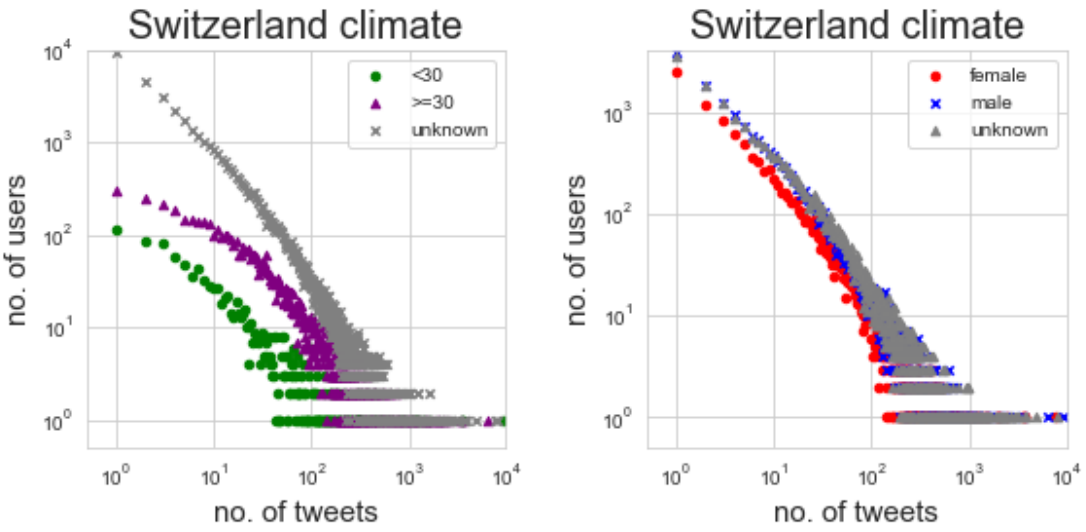

Figure 62. Distributions of users by number of tweets in Swiss about climate change considering age range (left) and gender (right)

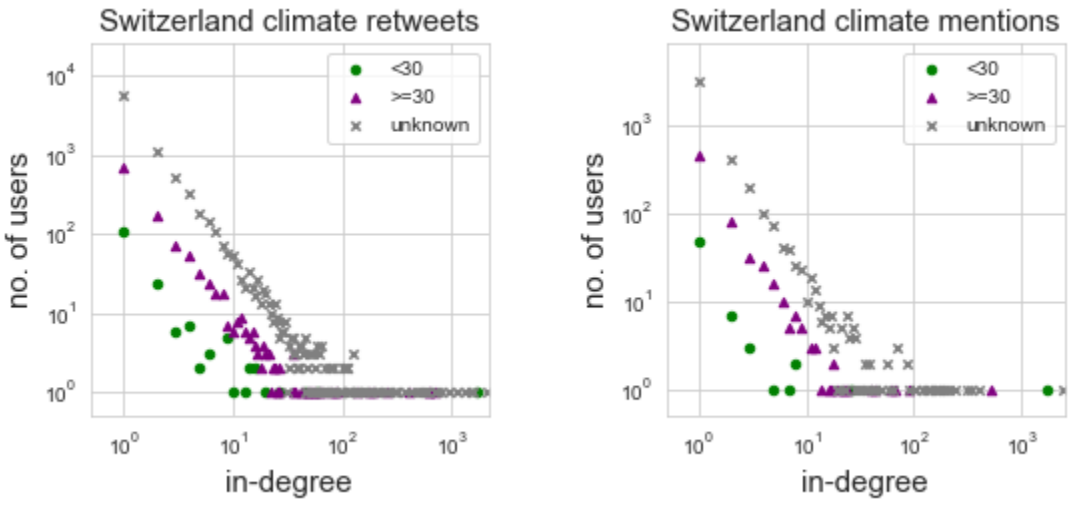

Figure 63. Indegree distribution by age in the Swiss networks for Climate change based on retweets (left) and mentions (right).

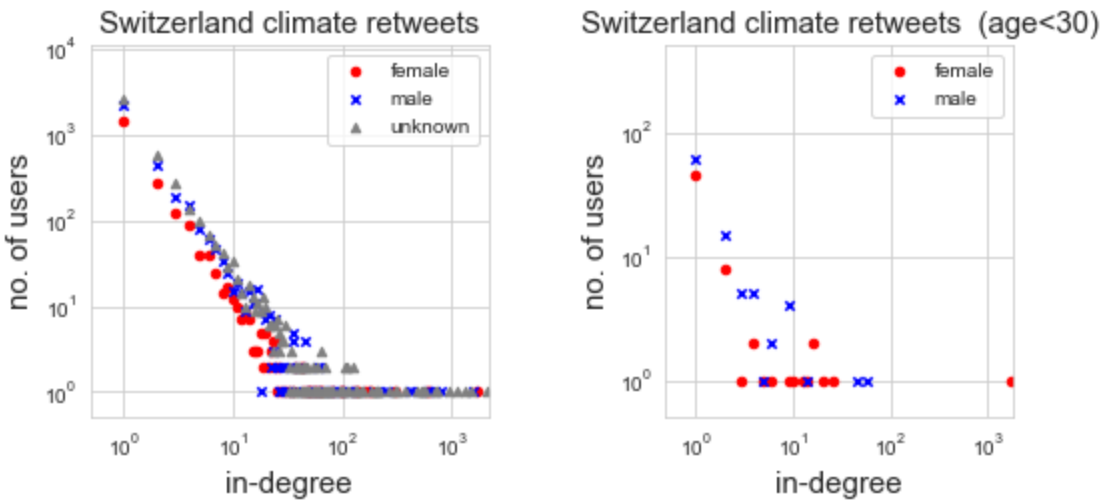

Figure 64. Indegree distribution by gender in the Swiss networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

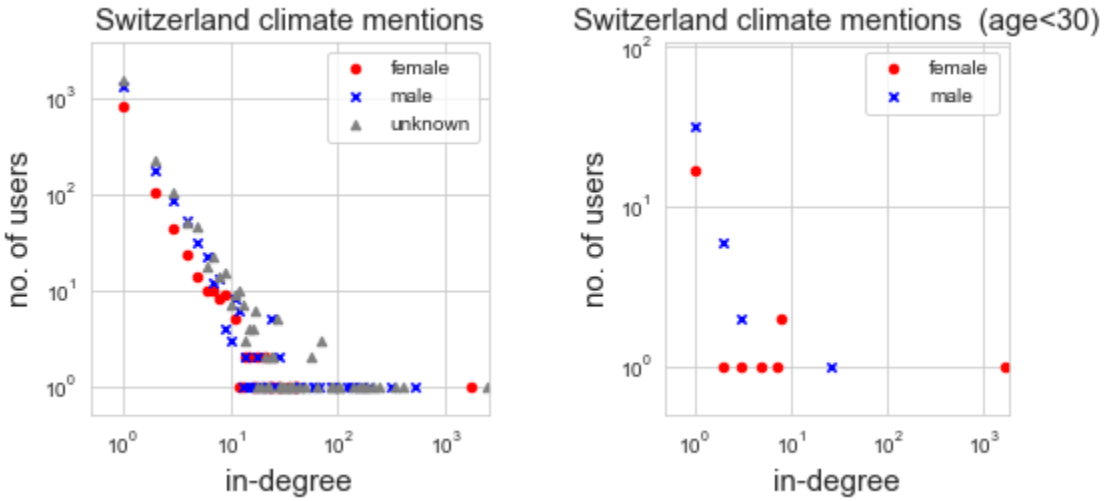

Figure 65. Indegree distribution by gender in the Swiss networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.8.2. Feminism

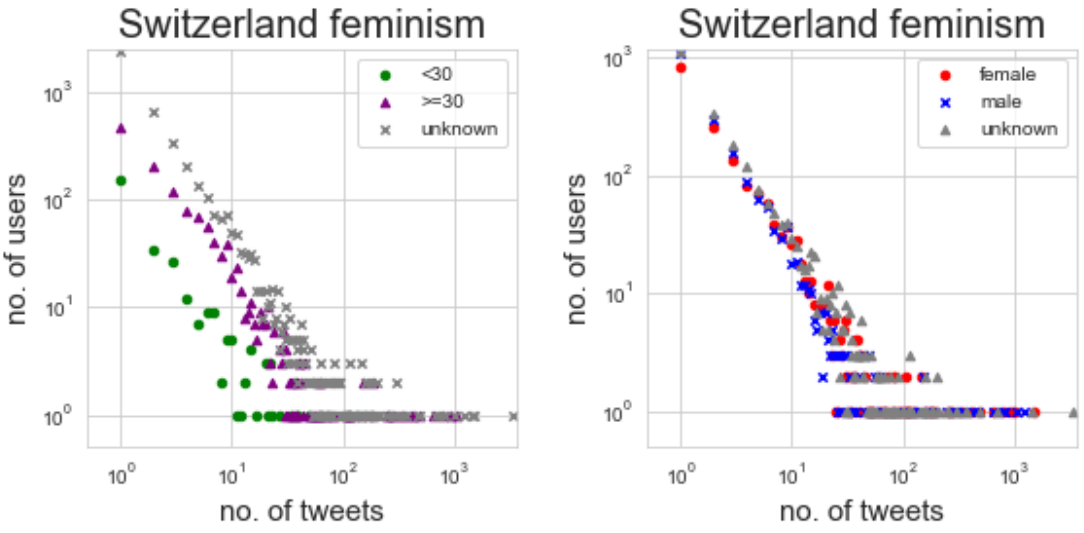

Figure 66. Distributions of users by number of tweets in Switzerland about feminism considering age range (left) and gender (right)

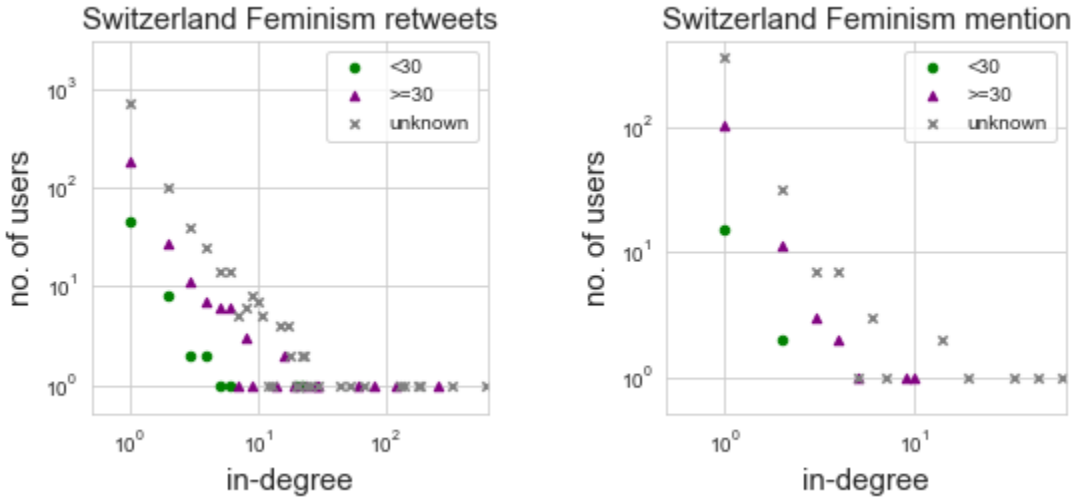

Figure 67. Indegree distribution by age in the Swiss networks for Feminism based on retweets (left) and mentions (right).

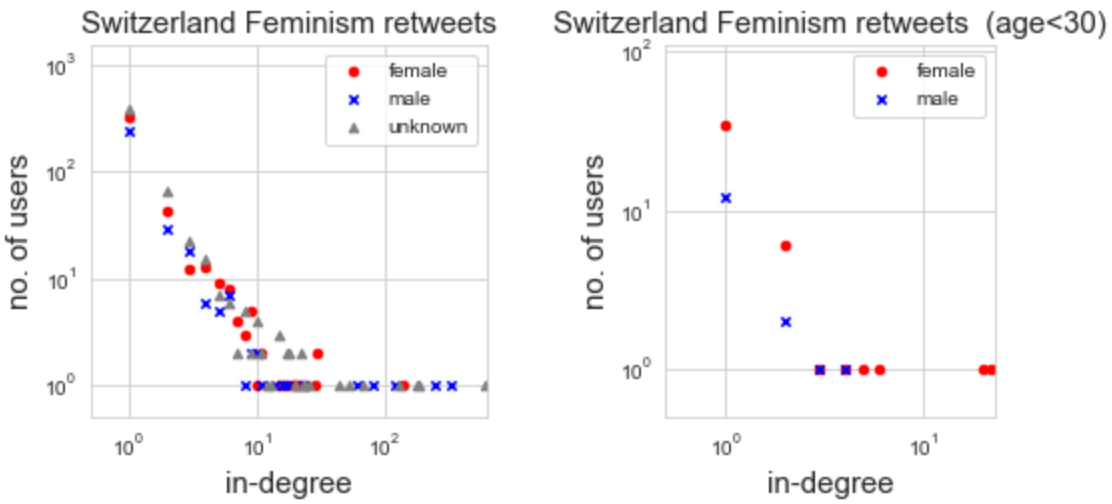

Figure 68. Indegree distribution by gender in the Swiss networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

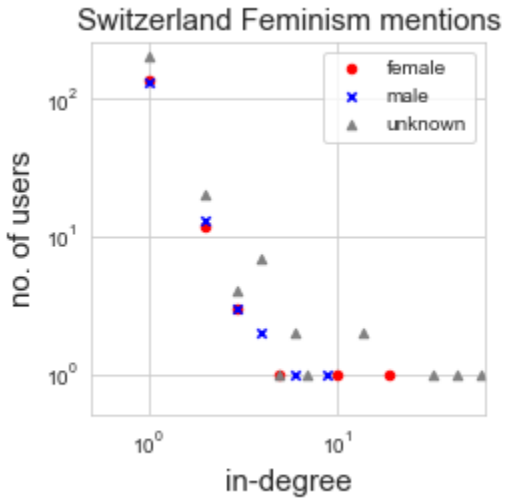

Figure 69. Indegree distribution by gender in the Swiss networks for Feminism based on retweets..

3.3.9. United Kingdom

3.3.9.1. ClimateStrike

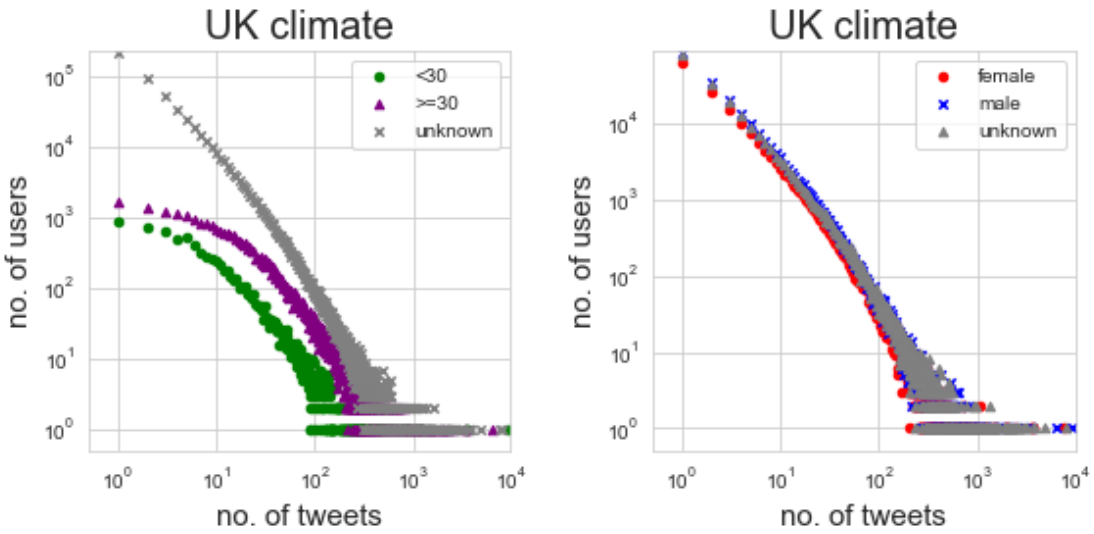

Figure 70. Distributions of users by number of tweets in the United Kingdom about climate change considering age range (left) and gender (right)

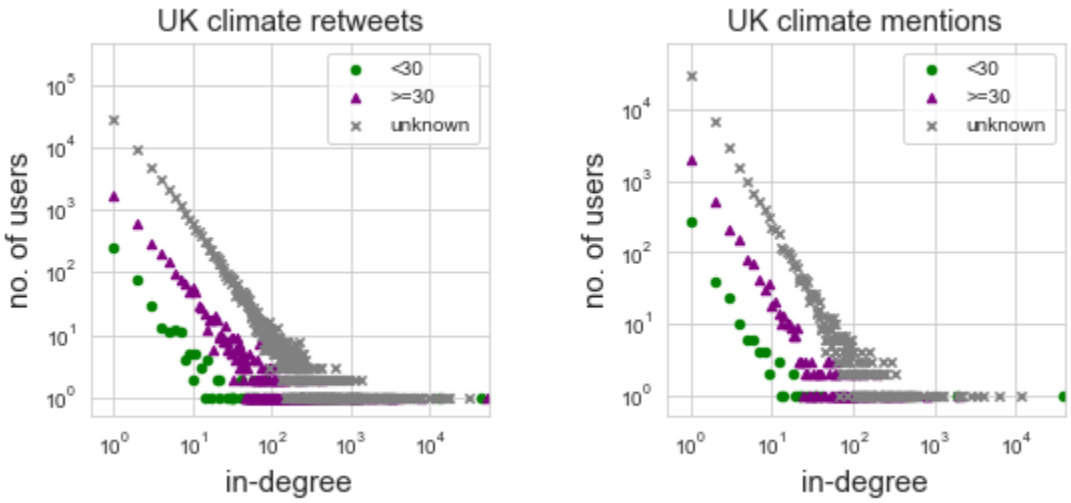

Figure 71. Indegree distribution by age in the UK networks for Climate change based on retweets (left) and mentions (right).

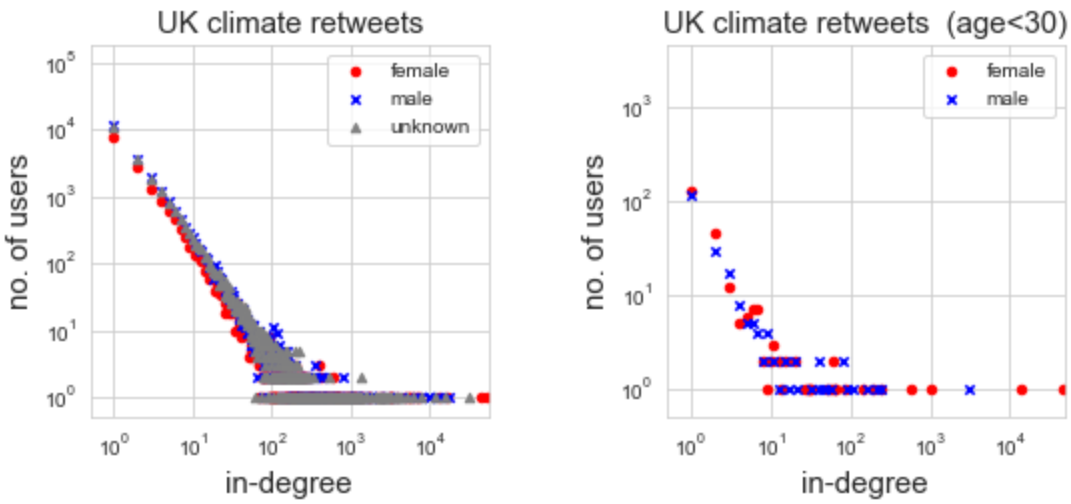

Figure 72. Indegree distribution by gender in the UK networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

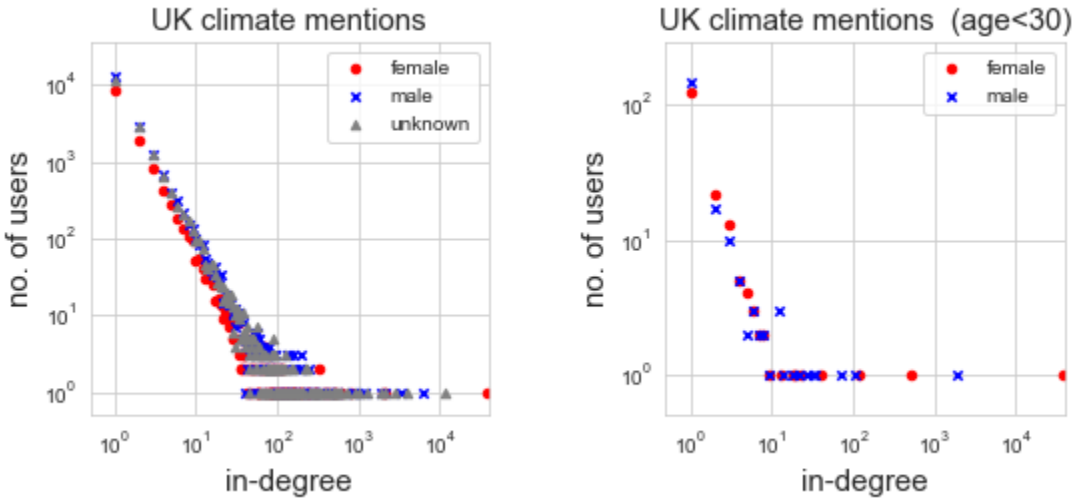

Figure 73. Indegree distribution by gender in the UK networks for Climate change based on retweets, for all users (left) and only for users below 30 years old (right).

3.3.9.2. Feminism

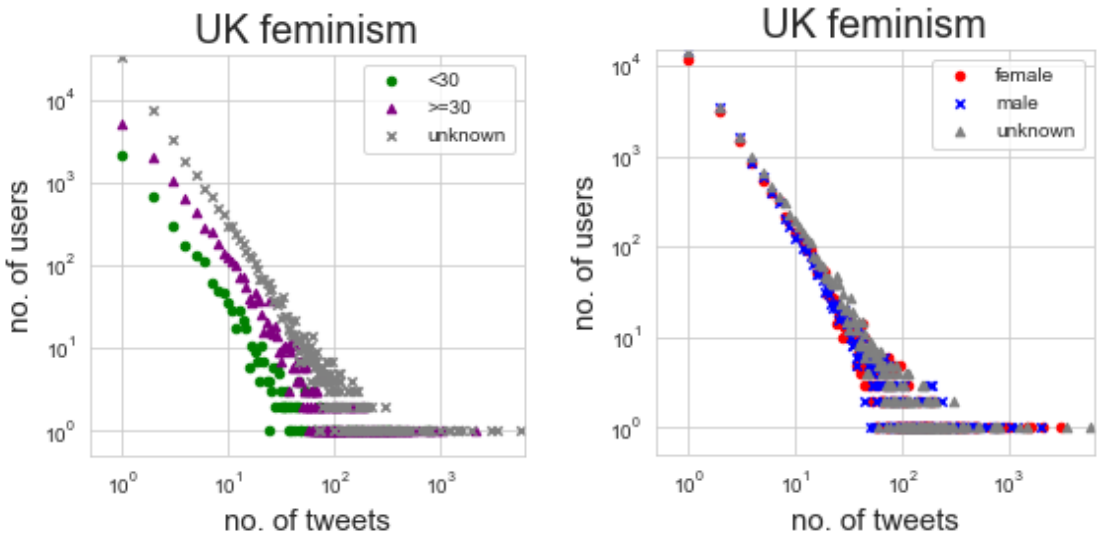

Figure 74. Distributions of users by number of tweets in the United Kingdom about feminism considering age range (left) and gender (right)

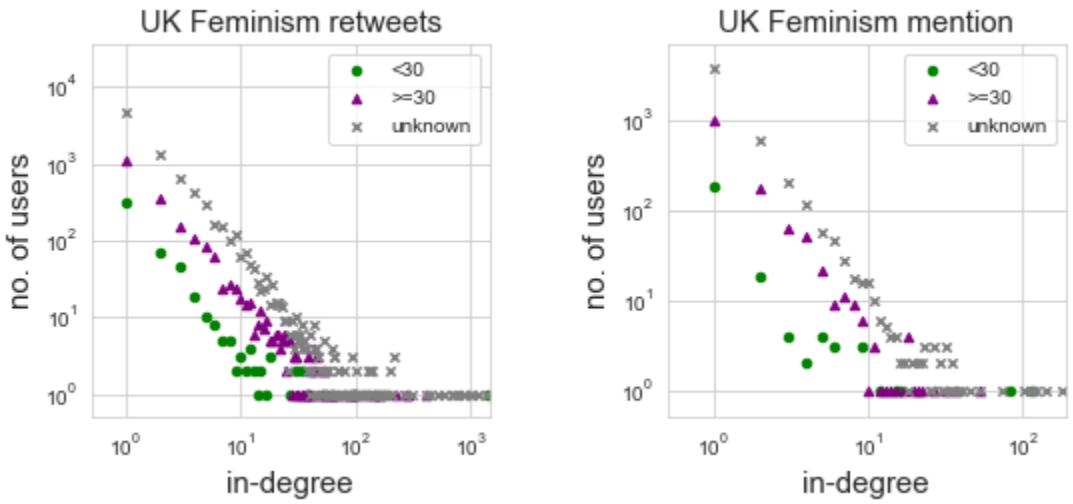

Figure 75. Indegree distribution by age in the UK networks for Feminism based on retweets (left) and mentions (right).

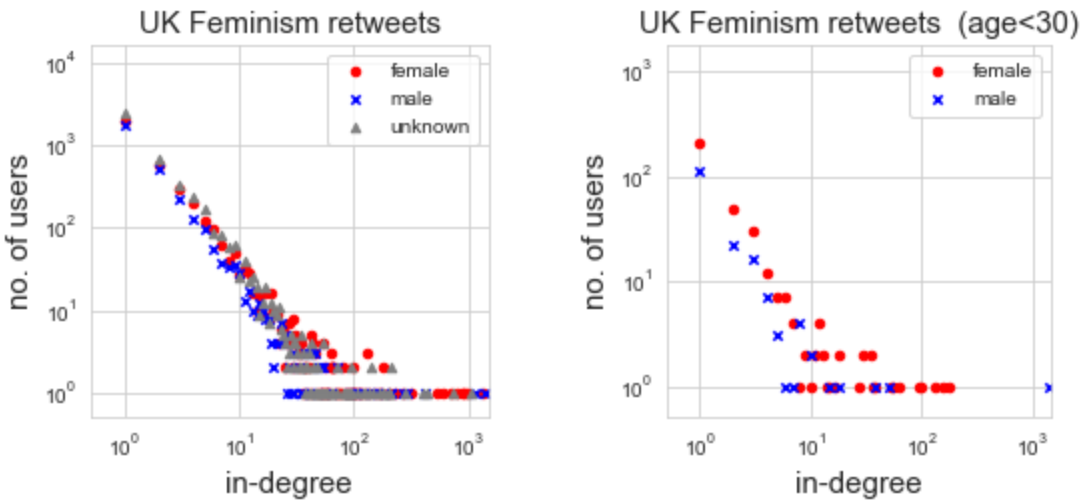

Figure 76. Indegree distribution by gender in the UK networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

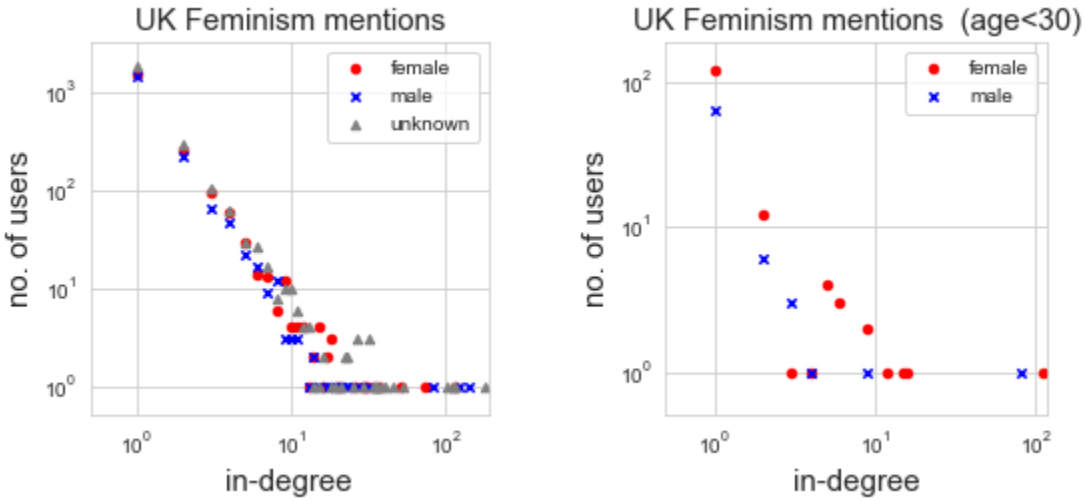

Figure 77. Indegree distribution by gender in the UK networks for Feminism based on retweets, for all users (left) and only for users below 30 years old (right).

4. CONCLUSIONS

The main value of the work presented in this document is that of proposing a data collection and analysis methodology for a cross-country study covering nine European countries, and creating datasets and results for each country.

This has implied a special effort for accounting for the intrinsic differences between the scenarios of the different countries, and the issues associated to special cases. We had to define special strategies to deal with countries that cannot be identified with a language, such as the UK, with English spoken at a global level, or Spain, with Spanish widely spoken in many Latin American countries, or Switzerland where various languages are spoken, overlapping with other countries. In these cases, the ability to detect the country from the user location indicated by the users was essential in order to filter messages and users by country. We also had issues with Greece, for which it was hard to retrieve a sufficient amount of data, and different criteria were combined to create the dataset.

Then, a critical point was that of developing a demographic analysis of inequalities, without any demographic metadata being explicitly associated to the users; the ability to infer demographic data for each Twitter account through state of the art methods was fundamental to allow for a deeper analysis of inequalities across countries, accounting for age and gender of the users.

As it could be expected according to previous literature, we observe a higher presence of men in the debates. This is generally true both in terms of number of users involved (as it can be seen in tables X-Y, reporting the homophily analysis results) and of activity and centrality levels. Interestingly, this is also the case for several countries in the debate about feminism, where one could naturally assume a major involvement of women. Women tend to be a minority, and tend to have a higher homophily, i.e. a higher preference for interaction with other women, higher than the preference of men for interacting with other men.

Two countries from Southern Europe present an exception: in Italy and Spain women have a comparable presence to men in the debate on Climate change, and are a majority in the debate on Feminism. Interestingly, in these countries women tend to have a neutral preference, i.e. no preference for interacting with other women, while in some cases men have a higher homophily, in the Spanish debate on Feminism where they are a minority.

The analysis of inequalities by gender unveils that men are not only a majority in most networks on Climate change, but tend also to be more active and central in these conversations, but in the cases of Spain and Italy. In the networks built for Feminism, instead, women are often less active in terms of number of tweets, but equally or more central than men in the networks of mentions and retweets; this is the case for most countries, and especially marked for Spain, where women's centrality overcome men's centrality by a big gap. These phenomena are even more marked when we restrict the analysis to users below 30 years old.

Beyond results for gender, we have presented also results for age difference; however, results in this case are less representative, as age range could be inferred with sufficient accuracy only for a minority of users, so the underlying patterns may remain in part uncaptured.

In the scope of the project it was possible to perform analyses and show results based on some relevant variables for all the countries. We chose to focus on two relevant demographic variables (namely age and gender) and two main metrics quantifying activity and centrality (namely the number of tweets and the in-degree in the interaction networks, respectively).

Further analyses could involve other variables: on the one hand, further metrics of individual relevance or centrality that were computed for each user, such as pagerank, outdegree or k-index in the interaction networks; on the other hand, further user attributes retrieved or inferred for each user, such as being an organization or not (as estimated through the m3inference library for inferring demographic information), seniority (based on the registration date or on the total number of tweets posted), influence in the social network, in terms of number of followers, growth in the number of followers during our observation period, geographic location.

All of these variables are included in the datasets generated with this document, and may be leveraged for extending the results presented here with further analyses.

We believe that a geographic analysis could be particularly relevant for assessing to what extent the debate within a country may be centralized in big cities, in urban areas, or in specific regions. This kind of analysis would be possible with the datasets we have produced, that includes a mapping of user self-reported locations to countries and cities, providing homogeneous locations.

5. REFERENCES

Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., & Vespignani, A. (2005). k-core decomposition: A tool for the visualization of large scale networks. arXiv preprint cs/0504107.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. Physical review E, 70(6), 066111.

Kleinberg, Jon (1999). "Authoritative sources in a hyperlinked environment" (PDF). Journal of the ACM. 46 (5): 604–632. CiteSeerX 10.1.1.54.8485. doi:10.1145/324133.324140.

Laniado, D., Volkovich, Y., Kappler, K., & Kaltenbrunner, A. (2016). Gender homophily in online dyadic and triadic relationships. *EPJ Data Science*, *5*(1), 19.

Napalkova, L., Aragón, P., & Robles, J. C. C. (2018). Big Data-driven Platform for Cross-Media Monitoring. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 392-399). IEEE.

Page, Lawrence, et al (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., & Jurgens, D. (2019, May). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In *The World Wide Web Conference* (pp. 2056-2067). ACM.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks. *Nature*, *393*(6684), 440.

Park J., Barabási A., (2007). Distribution of node characteristics in complex networks.Proceedings of the National Academy of Sciences 104, 46(2007), 17916–17920.