EURYKA

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

Guidelines for the Social Media Analysis

(Deliverable 7.1)

Workpackage 7: Social Media Analysis

Workpackage Leading Institution: UOC

Submission due date: March 2019

Actual submission date: June 2019

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

Table of Contents

1.OBJECTIVES OF THE STUDY 3

2. SAMPLING 4

2.1. Twitter sampling (partners) 5

2.2. Facebook sampling (partners) 7

3. DATA RETRIEVAL (EURECAT) 7

3.1 Extracting Twitter data 7

3.1.1 Inferring age 8

3.1.2 Automated gender identification 8

3.2 Extracting Facebook data 8

4. ANALYSIS OF INTERACTIONS BETWEEN TWITTER ACCOUNTS 9

4.1. Creation of interaction networks 9

4.2. Analysis of centrality 9

4.3. Structural characteristics of the networks of interactions 9

4.4. Communities analysis 9

4.5. Analysis of polarization and fragmentation 10

4.6. Analysis of inequalities 10

4.7. Homophily and cross-cutting interactions 10

5. FACEBOOK ANALYSIS 11

6. PRACTICAL ISSUES 12

6.1. Tasks WP7 12

6.2. Documentationand outputs 122

6.3. Questions and suggestions 122

7. BIBLIOGRAPHY 12**3**

1.OBJECTIVES OF THE STUDY

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 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 characterised by the idea that the young are "digital natives" – does not occur in a uniform or generalized manner, but rather happens 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 analyse how social inequalities are manifested in the way that young users actively participate (or do not) in politics.

The coordinated study (WP7) will be carried out in the nine countries participating in the project (France, Italy, UK, Germany, Poland, Greece, Spain, Sweden and Switzerland). Two analyses will be developed: one in relation to Facebook and the other in relation to Twitter. The objectives of each of these analyses are:

1) To study how youth-led and youth-oriented organizations provide opportunities to participate via Facebook, and how individual young people and organized networks use these opportunities across the nine countries.
2) To investigate young people's ways of doing politics online, and the impact inequalities has on this, by examining young people's use of Twitter. The goal is to see how young people in these nine different countries access and use Twitter for political purposes.

This analysis faces an important challenge related to the project's focus: there is no direct access to the personal data (e.g. age, gender, geographical location) of Twitter and Facebook users.

Given this challenge, the EURYKA Consortium has considered it necessary to outsource to a service highly specialized in the extraction and monitoring of social media data. Thus, outsourcing has been carried out to the services of the Eurecat research centre.[1]

Eurecat is a big industrial technology provider based in Barcelona. They offer applied R&D services, technology consulting, highly specialized development of innovative products, and promotion and dissemination of technological innovation. The subcontracted services are developed by the Data Science & Big Data Analytics Department. Particularly, subcontracted services are directly linked to their expert working areas regarding data mining, social media and computational social sciences. This outsourcing process will allow access to both the tools developed by this company and the latest advances in the fields of computer and data science.

[1] https://eurecat.org/en/

2. SAMPLING

The sample will be obtained manually by the research teams of each country. The scope of study has been limited to the debates generated on social networks around two themes: climate change and gender. Only two topics were selected in order to limit the magnitude of the search and allow the analysis to be feasible in the time allocated for this work package. The topics have been chosen based on the following criteria:

1. That they are current issues or debates in the public sphere;[2]

2. That these are issues to which the EU gives priority in its youth strategy (EU, 2018);

3. That both issues are global, with impact on a national-local scale.

Two search terms have been identified for each topic. The basic criterion for the selection of these two terms is that they are concepts with univocal meaning (that semantically do not present more than one meaning). Following this logic, the selected search term for climate change is *FridaysForFuture*, while the search term for the gender topic is *feminism*.

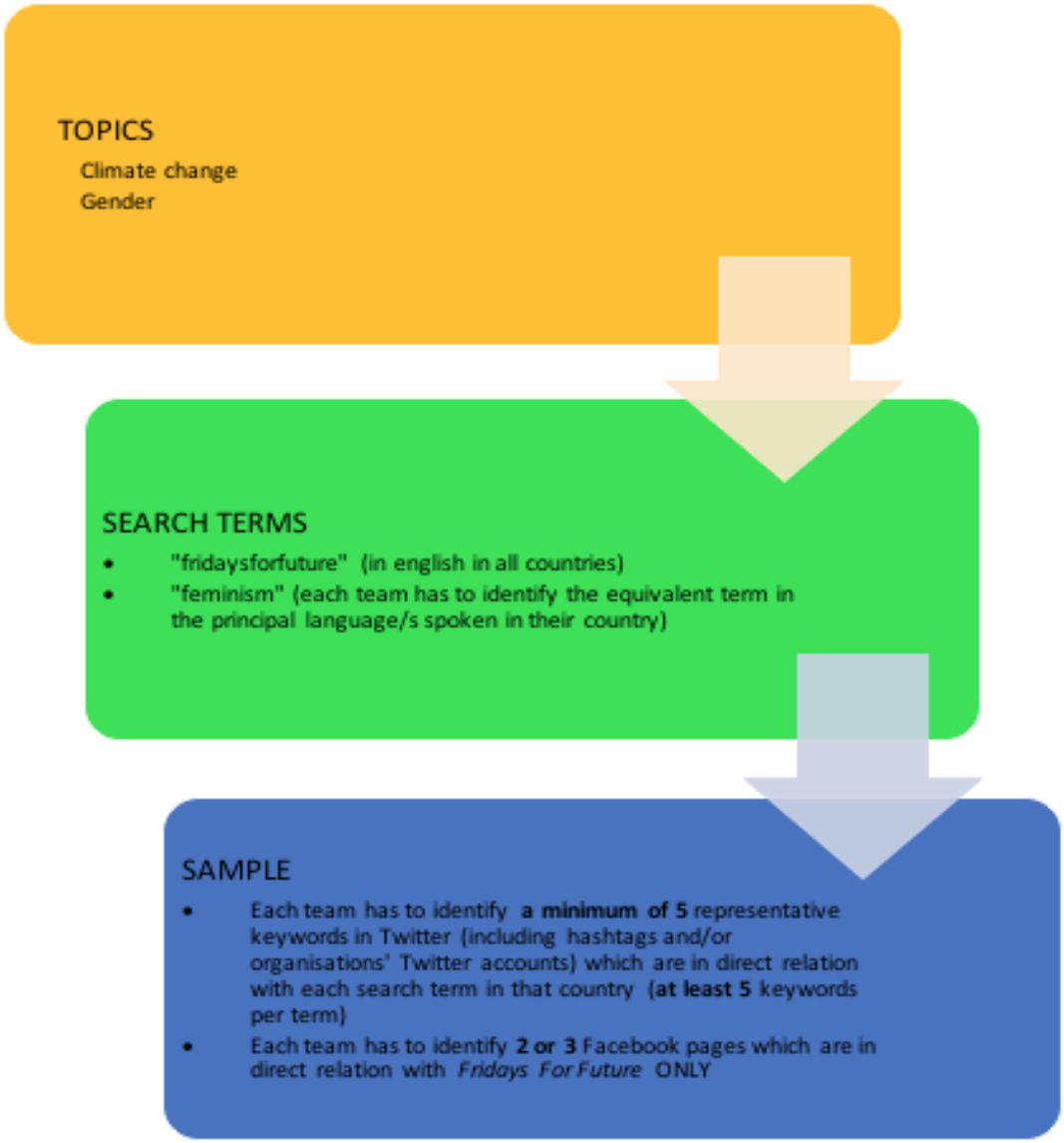
- **Fridays for Future**: A term currently widespread in the citizens' and media agenda. It refers to a movement led and mainly created by young people. It is considered a global youth campaign that can be analysed locally; dominant debate exists in English. Climate change is one of the goals of the EU Youth Strategy 2019-2027 (EU, 2018).

- **Feminism**: A topic that has different manifestations at national/local scale and with different capacities to transcend local and global debates. This is a subject about which a controversy and polarization of positions exist on different scales. Feminism is one of the goals of the EU Youth Strategy 2019-2027 (EU, 2018).

Thus, the analysis of young people's interactions on Twitter and Facebook will be based on a sample that will be obtained after the identification of representative keywords (in the case of Twitter) and representative Facebook pages (in the case of Facebook) in direct relation with the two search terms in the case of Twitter, and in direct relation to the search term "Fridays For Future" in the case of Facebook.

In the case of Facebook, the best option is to focus on "Fridays For Future" because it is an official Facebook account with twin accounts in different countries; identifying these accounts, and comparing them across countries should be straightforward. For feminism, there is no guarantee of the existence of any rigorous and replicable way of detecting accounts, as far as the accounts that really lead the movement can have different names in different countries.

[2] In the case of Twitter it is possible to collect data almost without restriction, but only in real time with the API Streaming. This is the reason why it is important to focus on current issues.



Below follows an explanation of how (based on the two search terms) all partners will obtain the keywords used for Twitter data monitoring, as well as the Facebook pages that will be used for Facebook data monitoring. The obtaining of this sample must be carried out manually and in a contextualized way (more details below on the guidelines that the partners must follow). EURECAT will be responsible for data retrieval from social networks, as will also be specified below.

2.1. Twitter sampling (partners)

To search for the most popular and relevant hashtags and organisations' accounts related to *FridaysForFuture* and *feminism* in each country, the teams should use Twitter. These are the steps:

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 carried out 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 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 UK will have to add location(s) in "Places". First try 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 five** 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. 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 teams 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 teams find 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).

2.2. Facebook sampling (partners)

The sampling of Facebook pages (from which EURECAT will carry out the data monitoring) will consist in identifying two or three Facebook pages that are in direct relation with the term “Fridays For Future”. Here there are the steps to follow:

1. Open your browser in incognito mode. This will prevent the search from being done based on algorithms generated by the researchers' personal accounts.
2. Open Facebook using the EURYKA project’s account. Details of this account (name and password) will be sent to all partners in a private message.[3]
3. Type “Fridays for Future” in the search bar and press enter.
4. Right below the search bar, select “Pages”. On the left, in “Categories”, select the last option “Cause or community”.
5. Scroll down the results and **select 3** of the most popular (high number of likes), relevant and active pages on Fridays for Future in your country.
6. Paste the URL links (e.g. https://www.facebook.com/fridaysforfuturebcn/) in the Excel file provided.

3. DATA RETRIEVAL (EURECAT)

Eurecat will extract data from the social networks Twitter and Facebook. In the case of Twitter, the gender and age range of the users will be inferred as additional demographic information. The plan is to use different tools for the different steps of the data retrieval process. By default, open source tools will be used for this purpose, although some proprietary tools will be also considered, for example Kalium, a tool developed by Eurecat.

3.1 Extracting Twitter data

Kalium will be used to gather data from Twitter. This is a tool developed by Eurecat that allows one to efficiently and flexibly manage the tracking of social network data in real time. Kalium ensures scalability and flexibility when it comes to recovering and managing social network data. It allows the retrieval of information from different social networks, including Twitter. You can monitor topics in real time by collecting data based on different criteria, such as all tweets that contain a particular hashtag or that mention or retweet a certain account. The tool also offers advanced functionalities for analysing and visualizing social interactions (Napalkova *et al,* 2018).

EURECAT will use Kalium mainly to retrieve information from the Twitter streaming API, monitoring the hashtags and specific accounts identified as relevant for each country. The data will be obtained in real time during the monitoring phase, then processed to clean and enrich them with additional demographic information, before carrying out the analysis.

[3] Using a newly-created Facebook account (associated with a mobile phone number that has not previously been used by any other Facebook account) would be the only reliable way of searching Facebook without a bias. Otherwise, things will appear by relevance according to the user's activity and social ties.

3.1.1 Inferring age

In recent months, a new tool has been developed to infer the age of Twitter users. This is called, M3Inference and was presented at the WWW '19 The World Wide Web Conference, the reference international congress on the Web, in the field of computer science (Wang et al., 2019). In addition to inferring data such as the age of the users, the tool is designed to work on the basis of recognizing a plurilingual reality, essential for studies of a supranational nature, such as that of EURYKA.

3.1.2 Automated gender identification

To add additional demographic data about users, tools will be used to automatically identify a user's gender based on their name, comparing it with registered names for a given country, such as SexMachine[4] and Genderize[5]. The precision of these methods usually exceeds 80% for European countries (Karimi et al, 2016). The M3Inference tool also allows us to infer the gender of the users of the accounts.

3.2 Extracting Facebook data

For this, EURECAT will use the Netvizz[6] tool, developed by the "Digital Methods Initiative" group of the University of Amsterdam (Rieder, 2013). Despite Facebook's growing restrictions in terms of data access, which severely limit the possibilities of independent analysis (Rieder, 2015), it is possible to collect some data from public pages, such as the interactions around the posts published by these public pages

The data is anonymised and thus the characteristics of the users will remain unknown. Only their interactions will be shown. Therefore, in this case, it is not possible to extract any demographic information. Neither will EURECAT be able to automatically identify the gender of the users based on their name, as in the case of Twitter, since the names of the users are not available either. Therefore, data from public pages will be collected based on user interactions with the content generated by the sample pages.

EURECAT will gather posts, and anonymized comments and interactions around the identified accounts so that it will be possible to quantify volume of activity, received attention or reactions.

[4] https://pypi.org/project/SexMachine/
[5] https://pypi.org/project/Genderize
[6] https://tools.digitalmethods.net/netvizz/facebook/netvizz/

4. ANALYSIS OF INTERACTIONS BETWEEN TWITTER ACCOUNTS

Kalium and M3Inference will be used for this task. All the tweets including the keywords (hashtags or Twitter accounts) in question can be collected in real time, provided it does not exceed the total threshold set by the platform (corresponding to 1% of the total number of tweets at that moment). In such a case, some messages will be lost, and only a sample of the tweets will be obtained.

Relationships between Twitter accounts will be analysed based on three types of actions offered by the platform: retweets, replies and mentions. The retweets will help us investigate the flow of information, the most influential and retweeted actors in the spread of the messages, as well as levels of polarization and fragmentation. The replies will be studied to investigate patterns of discussion and debate, controversies and polarization. The mentions will help us understand the conversation patterns and identify the actors that receive the most attention.

4.1. Creation of interaction networks

For the three types of actions, directed networks will be created, in which each interaction (each retweet, reply or mention) corresponds to an edge (connection) between two nodes that represent two users/accounts, and in which, according to an established convention, a incoming connection to a node represents received attention. If a user A writes a tweet and a user B retweets this tweet, a directed edge from B to A will be generated in the retweet network. Similarly, if a user A writes a tweet and a user B replies with a direct reply tweet, a directed edge from B to A will be generated in the reply network. If a user A mentions a user B, a directed edge from A to B will be generated in the mentions network.

4.2. Analysis of centrality

To study the centrality of users in the obtained networks, different centrality metrics will be used to detect different aspects, such as the degree, or number of connections of a user with other different users, the pagerank as an indicator of relevance to a directed network, and the betweenness (centrality of intermediation) as an indicator of the influence on controlling the flow of information and bringing different communities and sectors of the network.

4.3. Structural characteristics of the networks of interactions

Where relevant, the structural characteristics of the networks will also be studied, such as the diameter or the agglomeration coefficient (clustering coefficient) (Watts & Strogatz, 1998) to characterize conversation patterns and themes across the different countries.

4.4. Communities analysis

To identify densely connected groups of users representing several sub-portions of the network (communities or clusters), clustering algorithms will be used (Emmons *et al*, 2016), among which the Louvain Method, which is particularly effective (Blondel *et al*, 2011).

4.5. Analysis of polarization and fragmentation

Based on the idea that a retweet generally represents an endorsement, as shown in previous literature (Connover *et al*, 2011), retweets will be used as indicators of affinity between users. In case of conflicting political issues and debates with opposing sides, a clustering algorithm will be applied to the retweet network to identify groups of users corresponding to the different positions in the debate. Once the groups of users with different positions in the debate have been identified, the levels of fragmentation and polarization in the network will be calculated, measuring the tendency of the users to interact with like-minded users. For this purpose, mix coefficient (Newman, 2003) or assortativity (Foster *et al*, 2010) metrics will be used.

4.6. Analysis of inequalities

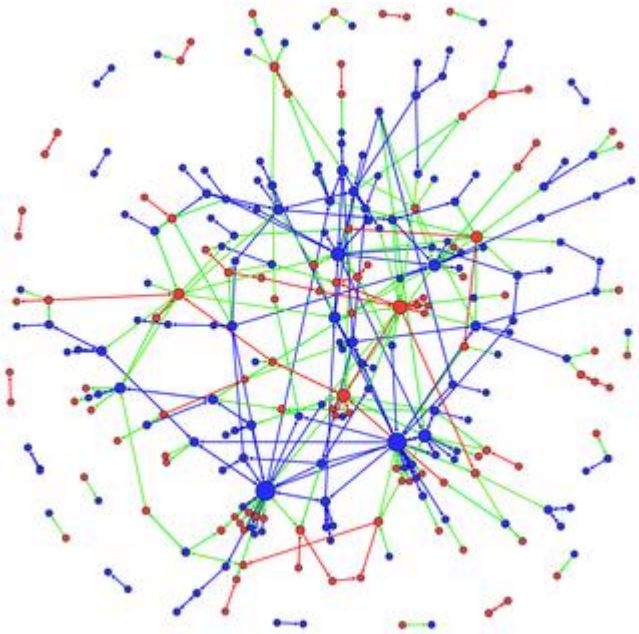
To study inequalities between different types of users, we will look at the distribution of basic metrics such as the number of tweets made, of retweets and mentions received, as well as the distribution of centrality metrics described above.

Then the available individual characteristics of the users, such as gender, will be crossed to investigate if they have an impact; for example, if women tend to be more active in discussing certain issues, or if they tend to receive less attention in terms of retweets or mentions, or to have less influence in the network according to the page rank.

4.7. Homophily and cross-cutting interactions

As a next step, the demographic characteristics of the users will be used to study how different types of users with different attributes relate to each other. For example, if there is a preference for men to interact with other men, or for women to interact with other women. This will be measured by estimating mix coefficient (Newman, 2003) and assortativity metrics (Foster *et al*, 2010), applied to the *a priori* characteristics of the nodes.

As a visual example, the figure below shows the network of interactions between Wikipedians that support the Republican Party and the Democratic Party of the United States. It can be induced visually that this network has a neutral mixing coefficient, that is to say that there is no preference for interactions between people who support the same party, nor for mixed interactions. This happens in Wikipedia, but it does not usually happen in social networks, where homophily is usually observed not only according to political ideology but according to many other attributes such as gender, age, origin, profession, etc.



Reply network in Wikipedia. Blue nodes represent Democrats, and red nodes Republicans. Green links represent mixed interactions, i.e. replies between a Republican and a Democrat editor. (Source: Neff et al,. 2013).

5. FACEBOOK ANALYSIS

Given the severe restrictions to data collection on Facebook, the analysis will be limited in this case to quantifying the volume of activity around the selected public pages from each country, relying on metrics such as the number of comments and reactions around the posts generated by a public page during the last month.

6. PRACTICAL ISSUES

6.1. Tasks WP7

Schedule of tasks required of each team:

Deadline	Task	Team	Guidelines page
June 20th 2019	(D7.1 submission)	UOC & Eurecat	

June 30th 2019	Twitter and Facebook sampling	All partners	4-7
July 2019	Twitter and Facebook data retrieval	Eurecat	7-9
September 30th 2019	5.1. Data analysis	UOC & Eurecat	9-11
October 2019	Provide report for each country (Facebook and Twitter)	UOC	
October 2019	Global report	UOC	

6.2. Documentation and outputs

By the end of the analysis, the Spanish team will deliver a small descriptive report for each country with visualization of the data analysis, as well as the global report with the transnational data analysis based on these descriptive reports (deliverable D7.2).

All documents will be sent by email to all partners.

Bibliography will be managed with Mendeley.

6.3. Questions and suggestions

Emails concerning WP7: Please draft emails to acluai@uoc.edu with the subject: WP7 + the specific issue that the email wants to address.

7. BIBLIOGRAPHY

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Proceedings of ICWSM*, 8, 361-362.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.

Conover, M. D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.

Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PloS one*, 11(7), e0159161.

European Union (2018). Resolution of the Council of the European Union and the Representatives of the Governments of the Member States meeting within the Council on a framework for European cooperation in the youth field: The European Union Youth Strategy 2019-2027 (2018/C 456/01), 18th of December https://ec.europa.eu/youth/policy/youth-strategy

Foster, J. G., Foster, D. V., Grassberger, P., & Paczuski, M. (2010). Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences*, *107*(24), 10815-10820.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6), e98679.

Jacomy, M., Girard, P., Ooghe-Tabanou, B., & Venturini, T. (2016). Hyphe, a curation-oriented approach to web crawling for the social sciences. In *Tenth International AAAI Conference on Web and Social Media*.

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 53-54).

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.

Neff, J. J., Laniado, D., Kappler, K. E., Volkovich, Y., Aragón, P., & Kaltenbrunner, A. (2013). Jointly they edit: Examining the impact of community identification on political interaction in wikipedia. PloS one, 8(4), e60584.

Newman, M. E. (2003). Mixing patterns in networks. *Physical Review E*, *67*(2), 026126.

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

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. In *Proceedings of the 5th annual ACM web science conference* (pp. 346-355). ACM.

Rieder, B. (2015). the end of Netvizz (?). *The Politics of Systems*. http://thepoliticsofsystems.net/2015/01/the-end-of-netvizz/

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.

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