Study of various forms polarization on the Reddit network

# Introduction

In this experiment we study various forms of polarization on Reddit. Reddit is a forum platform that people discuss any topic of their desire. It is divided into subreddits also known as communities. Each user can be part of any community that he likes. Each community has a set of posts or submissions - which is the same thing - that the users create and start a discussion.

-First we study ..

-Secondly ..

# **Pipeline preprocessing**

# General

The process selection of the data follows a specific pipeline. For each step, I save the data on the disk, so I can separate the steps of the pipeline. First, I need to select the appropriate subreddits. Secondly, for each subreddit I will divide the data in three categories. *Top, Controversial, Both.* For each category, I select a set of submissions and for each submission I create a comment tree. Then out of these trees, for each category, I create a user graph that represents the communication between users. All of this will be explained in detail below. The tool that has been used for the data collection is a module of python named *PRAW*, a wrapper library that handles the reddit API more conveniently*.*

# Subreddit selection

First, I made a general search on reddit with the related keywords and used the reddit filter ‘top’ for the resulted submissions. Then, I tracked the frequency of the corresponding subreddits for each resulted post. In other words, each post belongs to a subreddit, and I counted how many times each subreddit appeared in our search, so I can continue my research with the most famous ones. For the topic of covid vaccines I used the keywords ‘vaccin’ and ‘covid’ and for the topic of patriarchy I used the keyword ‘patriarchy’ and I ended up I the following subreddits

Covid Vaccines:

* r/Coronavirus
* r/conspiracy
* r/science

Patriarchy:

* r/WitchesVsPatriarchy
* r/MensRights

# Comment Tree Creation

Each post can be represented by a tree. The root of the tree is the post’s question/text and the branches are the answers of the of the post. So, each answer of a comment is a child of a node. If the comment is the answer to the comment , it means that the is the child of u and it is represented as: in the tree. For the tree representation I use the module *treelib* of python.

After choosing the subreddits, I crawled on each one and processed them on a specific way. For each subreddit I divide the data in three categories. *Top, Controversial, Both.* For the *Top* case, I put the filter *Top* that the reddit API provides and I collected the top submissions with the keyword ‘*vaccin’.* The same I did with the *Controversial* category. *Both* category is the union of *Top* and *Controversial*. For each category I create a set of trees, in the way explained above and saved them to the disk in the form of json. For each node, I save the comment data and the user that posted it.

# User Communication Graph

The next step is taking the trees and out of these create the communication graph. For each category (‘top’, ‘controversial’, ‘both’) I create a graph. I used two approaches. The first one is taking every edge of each tree, but now the node represents a user