Report Project Social Network Analysis

Bettini Chiara, Lestini Cinzia, Manassero Chiara, Spina Paolo

April - May 2021

1 Groundbreaking

At the beginning we have taught some possible argument to analyze. Each member of the group had explain to the others his/her chosen topic. After checking the feasibility or not of each topic, we have excluded some of them due to: lack of number of nodes, difficulty in finding and obtaining data. During this discussion some pieces of all ideas were catch and elaborated deeply. After all, we agreed in use Reddit as our environment for extract data. Nodes will be subreddits and for edges we decided to use crossposts. We have taken some time to understand better how Reddit works, have the time to obtain an API, understand which data were given to us and try to understand how scrape all data using python and praw library.

Our choice is to observe how subreddits, thanks to the crossposts, are linked and also which are their direction.

2 Data collection

Second step was elaborate different codes that allow us to crawl, scrape and clean all information, build a dataset and then a network. First, we create a code able to extract and scrape data using a Reddit API, starting from a chosen subreddit (environment). The major code (scrape-2.0.py) is based on the Benadith-first Search technique. It collects the fifity most popular posts present in the subreddit environment (level zero). From these posts the code picks which were posted in other subreddit, and the name of the subreddit were there is the crosspost is saved in a CSV file, in this way we found the first level. In a second time, the data collected are runned in another code (create-list-to scrape.py), allowing us to extact the name of new subreddits found. When a list of all these subreddits found is create, we run the new file in the major code. So we were able to find the second level and repeat it for the third one. At the end, we have obtain a csv file with all this information: starting subreddit of analysis; arrival subreddit; identification of that post; title; score; date; comments; if the post is original or a parent. At the end of this phase we build a code (clean-data.py) for delete redundant elements and possible parallel edges.

3 Network Characterization

Finished the data collection phase, we have used NetworkX for obtaining the major characteristic of our network and analyzing it. It is an oriented network, arrows follow the direction in this way: the original post is the source and the parent one is the target. Major features of our network:

nodes	24819
edges	107094
average degree	8.63000
diameter	10
density	0.00017
average clustering	0.10243
n. component	1

As we can notice our network is a medium size network. It's nodes degree on average is 8.63000. It is important underline that our network is a directed one so, if we calculate the average in-degree and out-degree, results are 4.3150 for both. On average a node have more or less 8 connections with the others nodes: 4 as target one and 4 as source. But, if we observe the density and the average clustering, it emerges that each nodes doesn't have so much connection with the others but most of them probably are connected to an hub. In one hand our major hubs sources, are the subreddits called: interestingasfuck, nextfukinglevel, funny with more the 2000 post that are crossposted on other subreddit. On the other hand, more targeted subreddits are: GoodRisingTweets, LateStageCapitalism, aww with 200 posts taken from other subreddit.

Our formed network is composed by one giant component, that is due to the costruction of our network. In fact our code is based on the Benadith- first Search technique, so is it impossible trace nodes that are not linked to our starting nodes.

For having a better comprehension of our network we have compared it with Random Network Model, Watts-Strogatz model and Scale free network model.

3.1 Random and configuration Model

For compare the model was fundamental create a Random Network Graph that has the same amount of nodes and edges of our subreddits graph. Starting the formula $L = p\frac{N(N-1)}{2}$, where L is the average number of links, N the number of nodes in the graph and p the probability of forming an edge, i computed the probability p necessary to have an ER graph with L equal to the number of links in the subreddits graph. Then the ER and subreddith graph degree distributions were computed in the subsequent graph:

(insert image of the degree distibution of two graph)

It is clear that our subreddit network follows a polinomial distibution $(a * x^b)$. On contrary the ER model follows the poissonian one.

Also all the other general feature were calculated on NetworkX for having a better picture of the other characteristics and what emerges was:

	ER	Configurational
weak conn comp	1	8
avg clust coeff	0.00035	0.03446
density	0.00035	0.00017
k max	38	2774
k min	4	1
k	9	same as real
avg shortest path	4.93	0.7
distribution	poissonian	same as real

4 Conclusion

References