A Reddit analysis

Bettini Chiara
c.bettini3@studenti.unipi.it
Student ID: 518134

Lestini Cinzia
21980920@studenti.unipi.it
Student ID: 219809

Manassero Chiara
c.manassero@studenti.unipi.it
Student ID: 547922

Spina Paolo
p.spina4@studenti.unipi.it
Student ID: 568773

ABSTRACT

(va compilato alla fine) In this network analysis we have tried to understand how are linked the different kind of subbreddits, following the crosspost activity between subreddits and further which are the subreddits more influent in our network, why some old post are reposted sometimes and how in a social, how some ideas could became so viral then were able to affect the financial market.
[1]

KEYWORDS

Social Network Analysis, Reddit

ACM Reference Format:
Bettini Chiara, Lestini Cinzia, Manassero Chiara, and Spina Paolo. 2021. A Reddit analysis . In *Social Network Analysis '21.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

(rigurdare una volta finito) In this paper we explore and comment all process of our network building starting form data collection, following the network characterization, searching the basic feature and analyze it with some major model (Random Network Model, Configurational Model, Watt-Storgraz Model and Scale Free Network). Then we explore further our network following these 3 questions: which are the major content spread on our network and is possible for them became viral? Why some old posts are crossposted also after one two year later? How is possible on a social like Reddit, that a community agreed to manipulate the financial market?

2 DATA COLLECTION

In this section we present the process that we have follow for decide Reddit as our source of data information, a simple explanation of our codes that: crawl, scrape and clean all data available using Reddit's API.

Selected Data Sources

Each member of the group have proposed a different topic and data sources environment in which was possible obtain data. After checking the feasibility or not of some different topic, excluding some of them due to: lack of number of nodes, impossibility to reach information, we agreed in use Reddit as our environment for extract data and try to observe and study his community users due to different reasons. First, it is unusual in Europe use it, his structure and use differs in comparison with others most common social network such as Facebook, Instagram, Twitter and so on. Second, few months ago, it was under the medias attention because a community decide to modify the financial market after have agreed it on Reddit. In addiction two of our member's group already use this social. Furthermore we were able to obtain Reddit's API quickly and using python's praw library for extract, collect and clean all data.

So our choice was to observe how subreddits, thanks to the crossposts, are linked and which direction follows this crosspost.

- Reddit as source of data
- subreddits as nodes
- crossposts as links

Crawling Methodology and Assumptions. The major code created (scrape-2.0.py) is able to extract and scrape data using a Reddit API. It is based on the Benadith-first Search technique. Starting from a chosen subreddit (environment), the code collects the fifity most popular posts present in the subreddit environment (level zero). From these posts the code picks and save, in a CSV file, all this information:

[1] **Project Repositories**
Data Collection: https://github.com/sna-unipi/data-collection
Analytical Tasks: https://github.com/sna-unipi/analytical-tasks
Report: https://github.com/sna-unipi/project-report

SNA '21, 2020/21, University of Pisa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$0.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(1) from (starting subreddit)
(2) to (subreddit were the crosspost appear)
(3) id (unique identification code)
(4) title (post's title)
(5) score (likes obtained)
(6) date
(7) comments
(8) parent (which subreddit have the original post)

All this information allow us to found the first level of interaction and collect some information that we could use further in the analysis of our network. In a second time, the data collected are runned in another code (create-list-to scrape.py), that were done to extact the name of new subreddits found, and create a list. When it was done, we ran the new file in the major code. So we were able to find the second level and repeat all these two passages for discover the third one. At the end, we built a code (clean-data.py) for delete redundant elements and possible parallel edges so we obtain a cleaned file.

3 NETWORK CHARACTERIZATION

For characterize our network we have used NetworkX library. Our network of observation is an oriented one. A node is a source if it have publish the original post. On contrary is the targeted if it has posted a parent copy of the original post. Also we were able to calculate all this following characteristics of our network:

Table 1: Characteristics of Subreddits Network

N	24819
L	107094
k	8.63000
d_{max}	10
density	0.00017
<C>	0.10243
component	1
<d>	3.736

*N means number of nodes, L links, k is the average degree, $d_{max} is the diameter, < C > represent the average clustering coefficient, < d > *[2] is the average path length*

As we can notice the network is a medium size one with 24819 nodes. on average nodes have degree 8.63, that means on average each node possess eight links. It is important underline that the average degree has to be divided between in-degree and out-degree because we observing a directed network. On average, our subreddits, have a node with in-degree of 4.3150 and out-degree of 4.3150. That means, on average, their major posts are cross-posted in 4 other subreddits and a subreddit posts 4 posts that are taken from other subreddit. But if we focus on density and average clustering, it emerges that each nodes doesn't have so much connection with the others, most of them are connected to an hub. Our major hubs sources are subreddits called: "interestingasfuck", "nextfukinglevel", "funny" with more then 2000 post that are cross-posted on other subreddits. On the other hand, subreddits that post more non original content are: "GoodRisingTweets", "LateStageCapitalism", "aww".

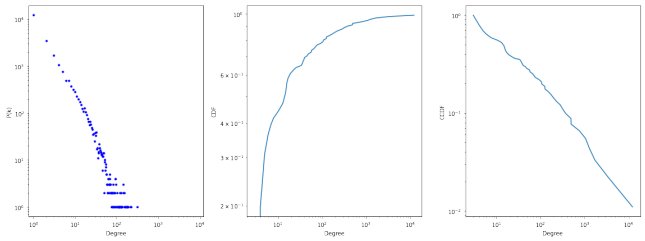

Figure 1: Degree distribution

Observing the first graph on right it emerges that we have a network with some few nodes heavily connected and lot nodes that posses just few connections.

According to the figure, it seem to follow a power-law distribution but due to the discrete nature of the degree distribution, for degree with small value the noise can be too high and it is difficult to understand if the distribution follows a power law. For this reason we used a logaritmic binning to better visualize the tail of the distribution. So under a certain value we divided data in a range of values and for each rage we pick the average value.

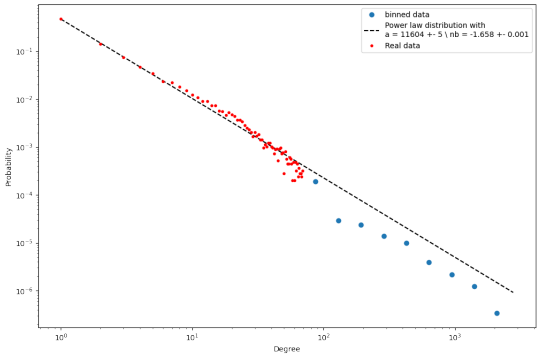

Figure 2: Binned degree distribution

From the figure it is clear that the distribution does not follow a power-low because the line from a certain point stay to much over the data binned data

Our formed network is composed by one giant component, that's because we have made a choice when we decided

to construct our network. In fact our code is based on the Breadth-first search technique: starting from a node, the code search his direct neighbors and then all neighbors of the neighbors until we reach, in our case, the third level. For this reason is impossible to us reaching nodes that aren't linked at least with one edges to our network.

This analysis was just the first phase, then we move on to compare our network with the following theoretical models: Erdos-Remyi Network, Configuration Model, Watts-Strogatz and Barbarasi-Albert.

Comparison with ER and Configuration model

modificare i dati che sono stati corretti

For compare our model was fundamental create a Random Network Graph that had the same amount of nodes and edges of our subreddits graph. Starting from $L = p\frac{N(N-1)}{2}$, where L is the average number of links, N the number of nodes in the graph and p the probability of forming an edge, we computed the probability p necessary to have an ER graph with L equal to the number of links in the subreddits' graph. So for obtain the probability we have used

$$p = \frac{L * 2}{N + (n-1)} \tag{1}$$

.

Then the ER and subreddits graph degree distributions were computed in the subsequent graph and we calculated all other characteristic to have a better picture of our network.

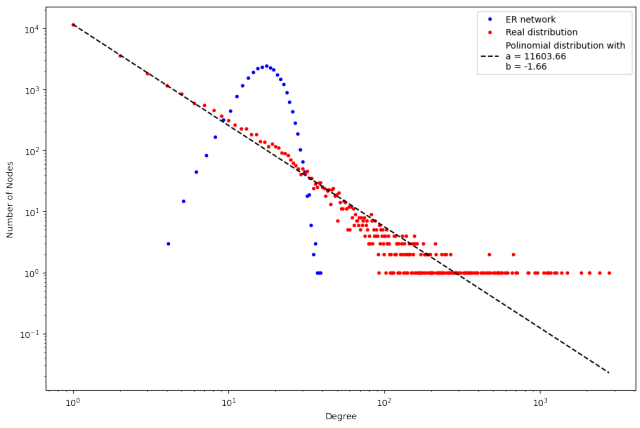

Figure 3: Degree distribution Subreddit network and ER
Subreddit network has a degree distribution that follows a polynomial distribution ($a * x^b$). On contrary the ER model follows the Poisson one.

From the comparison emerge that ER model represent a super-critical regime were $ln(N) > k = 9$ and have in common with our Subreddit network the average path length but absolutely is completely different if we observe the degree distribution. The configuration model could reach better our network in the degree distribution because it is built giving to each node a particular in and out degree. It also well represent the density due to the links creating following the degree of nodes, so degree distribution and density are connected one to each other. But the Configuration model fails in the representation of the average path length. In conclusion we could say that ER model Configuration model allow us to understand that our graph is situated in a super-critical regime with a big one giant component, that present loops and self loops, in addiction Configuration model reach something more in the degree distribution and density but for explore more is important comparing it with Watt-Strogatz Model.

Table 2: ER and Configuration models characteristics

model	ER	Configuration
wcc	1	8
<C>	0.00035	0.03446
dens	0.00035	0.00017
k.max	38	2774
k.min	4	1
<k>	9	same as real
<d>	4.93	0.7
distr	Poisson	same as real

wcc means weakly connected component,$< C >$ represent the average clustering coefficient, dens is the density of the network, k_{max} represent the highest degree distribution, k_{min} is the lowest one, $< k >$ is the average degree distribution, distr represent which degree distribution follow the model

Comparison with the Watt-Strogratz Model

Watts-Strogatz graph is a model for indirect networks with large clustering coefficient and short distances. As observed in real networks, it should be remarked immediately that our network (treated as an indirect one) has a not-so-high clustering coefficient (0.10243) but a low average short path length. First of all, we built a model, using three values: N nodes, K number of neighbours each link is linked to, and p probability of a link being rewired randomly to a different node (not allowing self-links). We treated our network as indirect one, so the average degree of nodes should be 8.404. Then we made two models with the same number of nodes of our network (24819), but setting the average degree k=8, for having the lower bound and k=9 such as upper bound. Doing this we obtain different number of links for each parameter k equal to 8 gave us 99276 links instead k=9 had 124095. As predictable no one of the two limits gave the same number of

our real network. For a Regular Lattice, Small World Regime and ER we calculated the major characteristics setting for each the upper and lower bounds.

Table 3: Table following Watts-Strogatz Model

	Lattice	ER	Small World
lower, upper	k = 8 | k = 9	k = 8 | k = 9	k = 8 | k = 9
p	0	1	0.37
<C>	0.6428 | 0.6667	0.0003 | 0.0004	0.1644 | 0.1695
dens	0.0003 | 0.0004	0.0003 | 0.0004	0.0003 | 0.0004
k	8 | 9	8 | 9	8 | 9
<d>	1551.625 | 1241.400	5.173 | 4.686	5.688 | 5.133
distr	Dirac delta func	poisson	poisson
connected	yes	yes	yes

Lower and upper refer to the boundaries, p is the probability to have a link with a distant node, <C> represent the average clustering coefficient, dens is the density of the network, k is the average degree distribution, distr represent which degree distribution follow the model, connected if all the nodes are connected

In one hand, Regular Lattice have an average clustering coefficient significantly higher then our network, in other hand ER have an average clustering coefficient lower then our. Small World Regime is near to our network in the average clustering and the average shortest path, characteristic that is also similar to our graph and ER model. Small World Graph is the model more similar to our Subreddit network but the density is different so we have to use another kind of network to explore this feature.

Comparison with Barbarasi-Albert Model

Scale-free networks are a type of network characterized by the presence of large hubs, that are a nodes highly connected to other nodes in the network. The presence of hubs will give the degree distribution a long tail, indicating the presence of nodes with a much higher degree than most other nodes. To understand if our network is Scale Free we have initially created an artificial Barabasi-Albert graphs, indirect. We create one of BA model putting the same number of nodes of our network and a number of links for each node equal to 8 (the average degree of our network if will be indirect). In figure 4 it's visible the BA model degree distribution, on the left, in the center the cumulative distribution function (CDF) and, on the right, the complementary cumulative distribution function (CCDF) or simply the tail distribution. Figure 5 is the computation of our network, considering it as indirect, using the same methods and graph. Figure 6 is the degree distribution of our real direct network. Then, Figure 7 represent the distribution of in and out degree in our network.

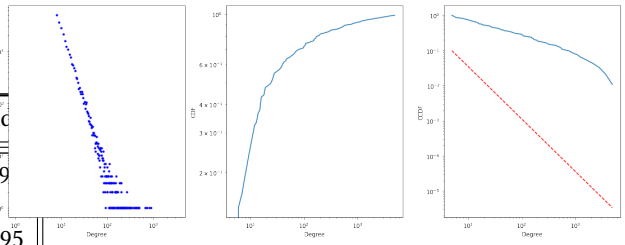

Figure 4: BA model degree distribution
The syntetich Network has alpha=2.505 and sigma=0.049.

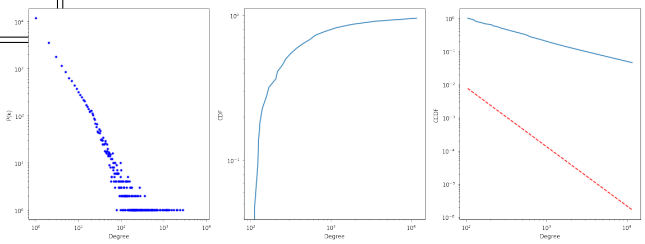

Figure 5: Subreddit network as indirect, degree distribution
Our network (transformed in indirect) has alpha=2.78 and sigma=0.168

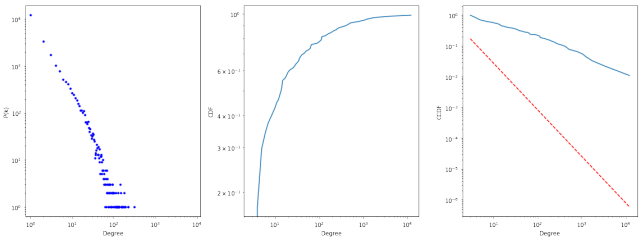

Figure 6: Subreddit network direct
The Directed Network has alpha=2.51 and, sigma=0.054

In conclusion, our network has a scale free network, since the alpha value falls within the range between 2 and 3. That's means in our network there is a presence of hubs as was already pointed out in the first part of the network characterisation.

4 TASK 1: COMMUNITY DISCOVERY ANCORA INCOMPLETA

In this section we decided to analyze the community discovery of our built network. Before starting we had to face one problem: our data are for direct graphs but if we use it with

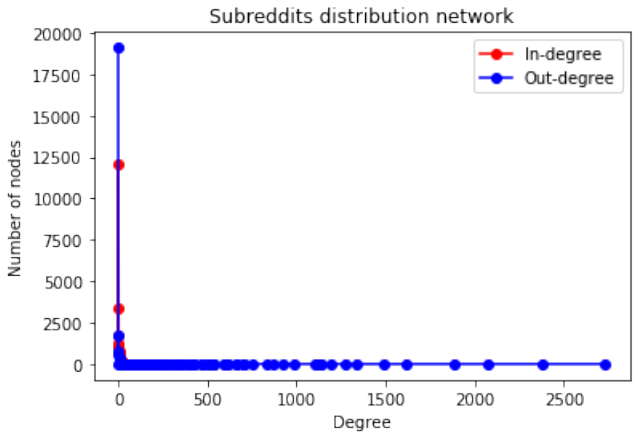

Figure 7: In ad out degree
As shown they are almost identical

NetworkX, we lose some attributes that are important, so we built a dataset suitable for indirect analysis without loosing information. What we have done was create two list that allow us to observe all links in both direction between nodes. Then, picking one of that list, inverting the name's columns "parent" to "to" and "to" to "parent", that's for maintain all the information about crosspost, interactions and so on, integrate them with the list not changed and melt in a single row all the information between each pair of nodes that have the same parent and to. That how we obtain a new data-set without loosing information. All the process could be seen in this .ipyn file "2.1.0)Build dataframeUndirectedGraph.ipynb"

Finally, having our new data-frame we could explore the presence of different community in our network. Using different methods present in the CDlib library, we try to understand which methods is more suitable for our network and then discover the community using it. The different discovery community algorithm used were: Laiden
Louviane Label Propagation Demon Infomap

For each algorithm we calculate clustering parameters, if ti overlap and the percentage of nodes that the clustering community covers.

algorithm	overlap	perc. node cover
Louvian	no	1.0
Laiden	no	1.0
Label Prop.	no	1.0
Demon	yes	0.3857
Infomap	no	1.0

Table 4: Algoritm and their output

Then we compute the fitness function index of clustering evaluation to obtain a synthetic representation of his minimum, maximum average and standard deviation values of their internal degree.

algorithm	min	max	score	std
Louvian	1.5	8.026	3.203	1.502
Laiden	1.0	8.121	3.012	1.552
Label Prop.				
Demon	1.5	8.026	3.203	1.502
infomap	2.388	8.512	3.934	2.643

Table 5: Fitness Function

Also we collect all the clustering evaluation making a comparison between the different algorithms, measuring their resemblance discovering the best matching between Louvian and Laiden algorithm, this is due because the first is a derivation from the second. Given all this clustering it was useful visualize how a given fitness function is distributed over the communities.

Figure 8: Community fitness/comparison visualization

Then we proceed making a qualitative evaluation analyzing the purity of each community, identifying which are the most homogeneous clusters, evaluating the modularity of Erods Reny, of newman Girvan and the conductance evaluation. At the end we watch their correlation that confirm Louvian and Laiden the most correlated.

Figure 9: Community fitness/comparison visualization

Finally we could use the Louvian algorithm because as was demonstrate the more suitable, to explore our network.
(explanation in what was found with louvian)

5 TASK 2: SPREADING DIFFUSION -CI SONO COSE DA MODIFICARE SIA NEI DATI, SIA DA CAPIRE COSA CONSIDERARE

What we analyse here is how the information,considered as infection, are able to spread around in our network. For doing so we have used not only our network but also all syntetic model that we have built before such as ER model, BA model, WS model.

We starting by assuming that a meme or a trend as already infected the first percent of our network, assume also it to be very easy to adopt a 0.2 threshold. Although we iterate the process for 100 times.

(insert threshold figure ln65)

With a 0.2 threshold and 1 perc. already infected, the cascade is complete with the ER and WS models. In our network and in the BA model, however, it dies out immediately. Supposing that the percentage of already infected is higher (a meme that is already very popular), at 10 percent, the situation changes:

(insert img until ln71)

The simulation shows that with a higher percentage of already infected nodes from the start, the cascade happens with the BA models but fails to fully happen in our network, becoming stationary, and thus showing the presence of a cluster with density 1-0.2 in our network.

Then we try to explore more the spreading using the SI model. Were we start an infection and stop when all nodes became infect. The iteration this time will be of 1500, with an infection rate of 0.001 and starting assuming that 1 percent of the network was alreay infected.

(insert img from ln104 to ln 119)

The speed of the spreading varies across the models: from faster to slower, BA, ER, WS, and our network.

So due to the fact that our network is the slower we want observe what could append in the SIS model were we set the 1 percent of nodes already infected, the infection rate equal to the previous one, a recovery rate of 0.005 and the lambda would be 2 because $lambda = \frac{0.01}{0.005}$ and iterating it for 600 times.

(insert img from ln 280 to ln 296)

For a second simulation, a situation where <1 and an endemic state is reached (< $\langle k \rangle$; $\langle k \rangle$ is 8.63 for our network):

perc. infected: 1 infection rate: 0.005 recovery rate: 0.01

Therefore, = 0.005/0.01 = 0.5, and 0.01 < 0.005*8.63.

(insert img ln 299, ln300, ln 266,67,69, ln270, 72,73)

Finally, a situation where a "disease-free" condition is reached. The condition to achieve this is having > $\langle k \rangle$.

The following parameters have been chosen:

perc. of infected: 1 percent infection rate: 0.001 recovery rate: 0.05 iteration 200 times

(insert img from ln313 to ln 324)

Then we use a SIR model were a simulation where > 1 (i.e. over the epidemic threshold):

perc. of infected: 0.01 infection rate: 0.01 recovery rate: 0.005 iteration: 1000 times

(insert img from ln341 to ln356)

Finally, a simulation where < 1 (i.e. over the epidemic threshold):

perc. of infected: 0.01 infection rate: 0.005 recovery rate: 0.01 iteration: 1000 times

(inset img until ln 375)

From the comparison between models, it appears that the spread of the "contagion" in the real network is limited in terms of infected nodes compared to all the other models.

6 TASK 3: LINK PREDICTION

In this section we discuss the expansion of our network's links between his nodes. Using a random model and an unsupervised approach, defining as a set of proximity measures unrelated to the particular network. In particular the predictor used were Random, Common Neighbors, Jaccard, Katz, Graph distance, Page Rank, Simrank. These predictor works different, in some case work using the neighborhood measures, in other what is considered is the distance, paths between nodes, furthermore some of take into account the similarity between two nodes. Before starting we have to split the data-set in two parts, a training and a test set. The split point was chosen 12th April 2021 because the training-test split correspond roughly to the 80-20 percent of the total data. Doing so, what happen was that the number of unique posts considered in total were 120747 and edges present in the training set were 159926 although in the test set were 39279. Then we deleted all the node in the training and test set that weren't adjacent at least of 3 nodes. In this way we eliminated all the subreddits that are not likely to interact one with each other. Lastly, we created a new graph that contains nodes present in both training and test set. At this point we have obtained a core with 3091 nodes and 102776 edges, that represent the most active subreddits, divided in 90097 that were present before the 12th April, and 8930 that were attached after that date. Obtained the core graph we started the prediction phase. it is important say that each predictor returns a score between two nodes u,v that represents how likely an edge (u,v) will be form in the future. In this table below, are presented the classifier and the accuracy of the prediction done using the edges that gave to us the highest score.

Classifier	Accuracy
Random	0.002576
Common Neighbours	0.036954
Jaccard	0.033147
Katz, beta: 0.05	0.039194
Katz, beta: 0.005	0.052744
Katz, beta: 0.0005	0.059574
Graph Distance	0.015342
Rooted page rank, alpha: 0.01	0.033931
Rooted page rank, alpha: 0.05	0.033819
Rooted page rank, alpha: 0.15	0.034602
Rooted page rank, alpha: 0.3	0.035498
Rooted page rank, alpha: 0.5	0.036618
Simrank	0.003471

Table 6: Link prediction accuracy

Even tough we achieved on average a performance better than a random predictor, overall the results are not good. At

most we achieved an accuracy of 5.96 percent. This is due to different reasons:

(1) The internet is unpredictable: New trends,memes and topic of discussions may arise at any moment without notice. We used posts in the training set that are too old: The majority of the crossposts are fairly new, in fact the test set is composed of posts at most two weeks old while the training set contains posts from 2017. Since the internet changes so quickly, relationship between subreddits that are this old, may lead to wrong results.

To test the second hypothesis we repeated the precedent steps considering only the posts posted from March 2021 onward. So now the posts considered were 62387, edges in the training set 68947, and in the test set 28366. The core that emerged from this have 2358 nodes, with edges in training set 38194 and in the test set 6516.

Classifier	Accuracy
Common Neighbours	0.056016
Jaccard	0.046194
Katz, beta: 0.05	0.055709
Katz, beta: 0.005	0.069521
Katz, beta: 0.0005	0.070135
Graph Distance	0.018109
Rooted page rank, alpha: 0.01	0.039134
Rooted page rank, alpha: 0.05	0.039441
Rooted page rank, alpha: 0.15	0.039441
Rooted page rank, alpha: 0.3	0.038674
Rooted page rank, alpha: 0.5	0.039288
Simrank	0.005525

Table 7: Link prediction considering just posts from March 2021 on.

It is clear that the results are a little bit better, but still too low. The last thing was try to assign a weight to each link based on the number of up-votes of the crosspost, following the idea that a crosspost with a very low number of up-votes does not represent a strong link between 2 subreddits.

The results just improve little for the predictor page rank but still insignificant the result, also this analysis was conducted not on all classifier because the neighborhood measures depends on nodes in common so it would be without sense consider them. For concluding this part, we obtain a scars result due to the unpredictability of the network were the information and content are constantly added and forgotten fast in internet.

7 TASK 4: OPEN QUESTIONS

Each day, on Reddit, people everywhere in the globe, participate and create new subreddits for discuss, share information and content on topic that have in common such as sports, events, animal, politics, art, economy, science... Sometimes one of them are able to have such a strong impact that produce some concrete action in ours everyday life. Observing Reddit and his user intercations attract in particular our attentions and now, that we have a better picture of the network built, we would like to deepen the research following these questions that have guided our research:
which are the subreddits more active in publishing and diffuse contents, what is the con typology and are they community related?
why some old posts are cross-posted in other subreddit also if some years are passed?
using as inspiration the Game Stop Case, we observe how viral posts could spread in our network and their impact on reality.

Classifier	Accuracy
Graph Distance	0.008594
Rooted page rank, alpha: 0.01	0.041283
Rooted page rank, alpha: 0.05	0.041743
Rooted page rank, alpha: 0.15	0.042204
Rooted page rank, alpha: 0.3	0.041897
Rooted page rank, alpha: 0.5	0.042204
Simrank	0.00245

Table 8: Link prediction considering up-votes and comments

For includes that condition we use the logarithm of comment + up-votes + 1 as weight

Subreddit characteristics and activities

For observing subreddits characteristics and activities, in particular the ones that publish most original content. So in our data-set, we choose one-hundred subreddit that post most original content and classified them according to their:

(1) number of members
  (a) very high = num members > 10.000.000
  (b) high = num members > 1.000.000
  (c) medium = num members > 100.000
  (d) low = num members > 10.000
  (e) very low = num members < 10.000
(2) type
  (a) News: Subreddits focused on sharing news (example r/environment)
  (b) Media share: Subreddits focused on sharing media like images, videos gifs etc (example r/memes)
  (c) Community: Subreddits composed of people sharing a common interest (example r/gaming)
(3) theme
  (a) Politics: (example r/Europe) Animals/nature: (example r/awww) Science: (example r/technology)

(b) Memes: (example r/funny)
(c) News: News subreddits that are not focused on a single topic (example r/news)
(d) Melting pot: There is not a main theme, but rather a mixture of all the themes I listed above (example r/BeAmazed)

Once the dataframe was built with this categorization, we analysing the distribution of various class type of the twenty-five subreddits that create most original posts. It emerges that excepting one all the other most active subreddit have high or very high number of members, usually the type of these subreddit are news or media share and the subreddits' theme are quite diverse. """(inseriamo gli istogrammi? se sì, tutti?)"""

The only subreddit with major number of original content present from the top twenty-five with 4k members was r/ForUnitedStates. For this reason we have taken a closer look because it could be a a SPAM or it is full of active people? It certainly has 1929 number of post crossposted to other subreddits and 405 original post created. Most of his content have just one comment and 9.29 up-votes on average, if we not consider the posts left without comment the average raise up to 3, so a little, and up-votes did not change practically because it is 9.53 the average. Analysing the most five followed original posts, this subreddit seems to be a typo of news and deals with left-wing topics. On first impact do to his small amount of comments and interactions seems to be a genuine subreddit but, for be more sure, we also observe his most popular posts crossposted.

""(inserire immagine forunitedstates?)""

Then comparing r/ForUnitedStates with all the other 25th subreddits in average interactions and type we could say this subreddit could have real people and genuine connections because it have, on average, less interaction then the others according also to his members' numeber. In addiction if we filter the results by subreddit type we can see that all the news subreddits generates a lower amount of interactions in respect to the media share subreddits, this is respected in r/ForUnitedStates too.

"""(inserire istogrammi)"""

Following now the one-hundred subreddits with most original content, still media share are by far the most popular. What's point out are the emergence of medium and low members' subreddits but most surprisingly is the presence of ten very small subreddits. Also the typo of subreddits give more space to the community so people that shared common interests such as gaming, sports... of course media share subreddits remain massive and there are not quite differences between themes as we observed in the first twenty-fifth subreddits.

If we observe all these feature in subreddits that are content reposer we could say the same things about the subreddit that on average share more crossposts then original ones? As the analysis above we started with the twenty-fifth more active subreddits that share not original content.

"""(inserire istogrammi)"""

Very low and high-medium number of members used to share not original content, most of them are media share typo but community typo are more then news subreddits. Themes of this subreddit are politics, melting pot and meme the majority but we could notice that there are no science subreddits but emerges personal one. Those are private subreddits created and operated by a single redditor, in which all posts are crossposts from other subreddits. The reason why those exist is because there is a limit of 1000 posts a single redditor can save and once this limit is reached a personal subreddit is created in order to continue to save posts.

""" (inserire istogrammi)"""

If we took a large picture, taking into account the top one-hundred reposter subreddits, what seems to be different is the presence in only one subreddit with number of member over 10M and the majority of them are in the range 100k-1M.This may be caused by the fact that those smaller subreddits have less active users and are at the same time more niche, so less original content are created.

As showing until now we could say the major part of original content are created in subreddits with and high number of members and then they finish in small subreddits. The major part of the subreddits' contents are media share, following news and discussion on topic common between members. In our network there isn't a dominant theme in fact melting pot represent a category of subreddits that share variable contents, as was predictable, but is quite common on on our network find posts related on politics and memes. So if there isn't a preferred theme that could have more success then other, how a post could became viral and be cross-posted?

For try to figure how this could happen, we were thinking that probably more up-votes a posts have and higher is the probability of this post to be saw and crossposted. So we choose the first one thousand viral posts (having the highest number of up-votes) and we check in how many other subreddits they were posted.

""" insert img """

Surprisingly the number of subreddits its very low, infact the majority of the posts were crossposted in only 1 other subreddit. We make another test using the top ten thousand viral posts for compare the result but once again the number of crosspost sill very low. We try to watch his distribution but no significant change was found.

"""(manca l'immagine della distribuzione)"""

According to our observations it's not so easy for a post became viral and be cross-posted in hundred and hundred subreddits, probably it's not enough have a lot of up-votes, in part that could due because Reddit have a particular structure and his users participate and follow only what they are interested on. Other motivation could be because a post need a sum of circumstances for having a chance to became viral. Furthermore it also possible that in Reddit there aren't BOTs that spam contents in different subreddits.

8 DISCUSSION

REFERENCES