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, 14 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 followed to 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 proposed a different topic, a data source 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 to extract data and try to observe and study his community due to different reasons. Firstly, is less common in Europe use it instead of Twitter, Facebook, Instagram and so on. Nevertheless it has an unusual structure, in fact, reddit is a space where other sub-groups called subreddits exists. These sub-groups are used by people to discuss a common shared interest including also videos, pictures and so on. For example, inside the subreddit r/sport there is only content related to sports, in addiction, alongside it there could be new subreddits that talk about a specific such as r/hockey. Secondly, few months ago, inside a financial subreddit people found an arbitrage opportunity in the stock market and bought GameStop stocks until the price reached 400 $. In addiction two of our group members already use this social. Furthermore reddit does not block access to his application programming interface (API from now on), so we were able to use their interface that lets our program to easily access reddit data. Lastly, python's library praw allow us to reorganize extraction, collection and clean all reddit's data.

So our choice was to observe how subreddits, thanks to the crossposts, are linked and which direction these crossposts follow .

- Reddit as source of data

[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

- 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 the Reddit API. It is based on the Benadith-first Search technique. Starting from a chosen subreddit (environment), the code collects the fifty most popular posts present in the subreddit environment (level zero). From these posts the code picks and saves, in a CSV file, all this information:

(1) from (starting subreddit)
(2) to (subreddit were the crosspost appears)
(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 allows us to find 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 is runned in another code (create-list-to scrape.py), that extracts the names of new subreddits found, and creates 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 to discover the third one. At the end, we built a code (clean-data.py) to delete redundant elements and possible parallel edges so we obtain a cleaned file.

3 NETWORK CHARACTERIZATION

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

	Network characteristics
Nodes	24819
Links	107094
<k>	8.63
d_{max}	10
Density	0.00017
<C>	0.102
Weak components	1
<d>	3.74

Table 1: Characteristics of Subreddits Network

As we can notice the network is a medium size one with 24819 nodes. On average nodes have a degree of 8.63, that means on average each node possess eight links. It is important to underline that the average degree is 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 to other subreddits. On the other hand, subreddits that post more non original content are: "GoodRisingTweets", "LateStageCapitalism", "aww".

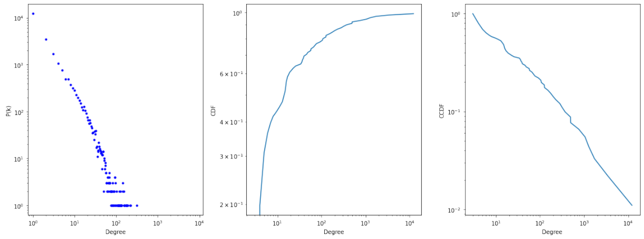

Figure 1: Degree distribution

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.

Our formed network is composed by one giant component, that's because we 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 we cannot reach nodes that aren't linked at least with one edge 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

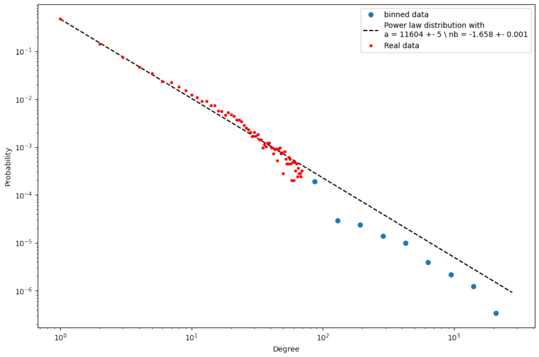

Figure 2: Binned degree distribution
From the figure it is clear that the distribution does not follow a power-low because the line from a certain point stay to much over the data binned data

To compare the various models was fundamental to 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 similar 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.

	ER	Configuration	Real data
Weak components	1	8	1
<C>	0.00016	0.03459	0.102
k$_{max}$	23	2774	2774
k$_{min}$	0	1	1
<k>	4	8.63	8.63
<d>	6.89	0.70	3.74
Distribution	Poisson	Power law	Power law

Table 2: ER and configuration model characteristics

From the comparison it emerges that ER model represents a super-critical regime where $ln(N) > k = 9$ and has 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 represents 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.

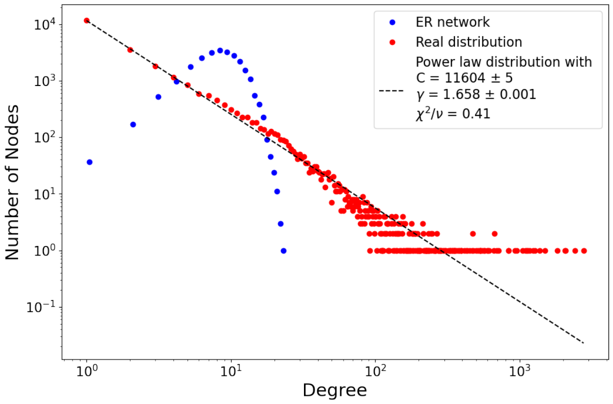

Figure 3: Degree distribution Subreddit network and ER

From the value of $\frac{\chi^2}{\nu}$ we can conclude that the degree distribution of our network follows a power law distribution. On contrary the ER model follows a Poisson one.

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
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

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.

In conclusion, our network is 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.

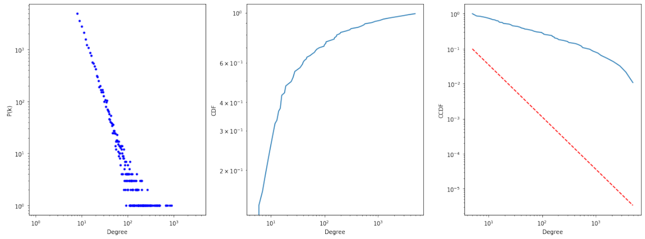

Figure 4: BA model degree distribution
The syntetich Network has alpha=2.505 and sigma=0.049.

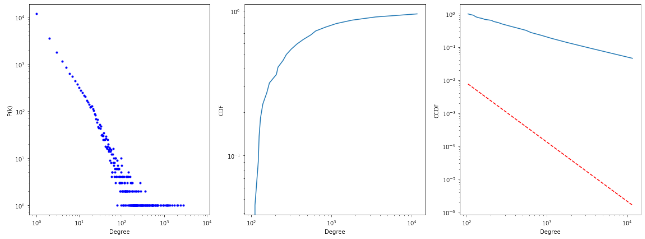

Figure 5: Subreddit network as indirect, degree distribution
Our network (transformed in indirect) has alpha=2.78 and sigma=0.168

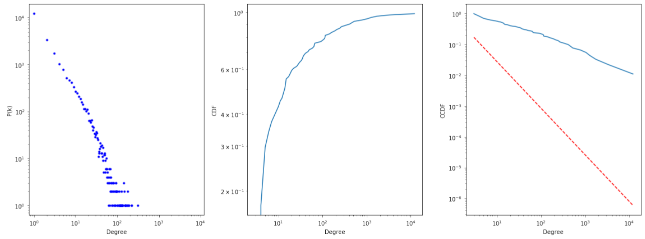

Figure 6: Subreddit network direct
The Directed Network has alpha=2.51 and, sigma=0.054

4 TASK 1: COMMUNITY DISCOVERY ANCORA INCOMPLETA

In this section we decided to analyze the community discovery of our built network. Using an algorithm to identify a meso-scale topology hidden within complex social network structure. Before starting, we have to face one problem: our data are for direct graphs but, if we use the "graph" function of NetworkX, we lose some attributes that are important. In order to avoid the loss of information, we built a data-set suitable for "graph" function that contain all the information. In particular, we have created two lists that allow us to observe in-links and out-links for each nodes. Then, picking one list, we inverted the name's columns "parent" to "to" and "to" to "parent". This switch, allow us to maintain all the information on cross-post, interactions and so on for each pair of nodes. What's more, we integrate the previous list

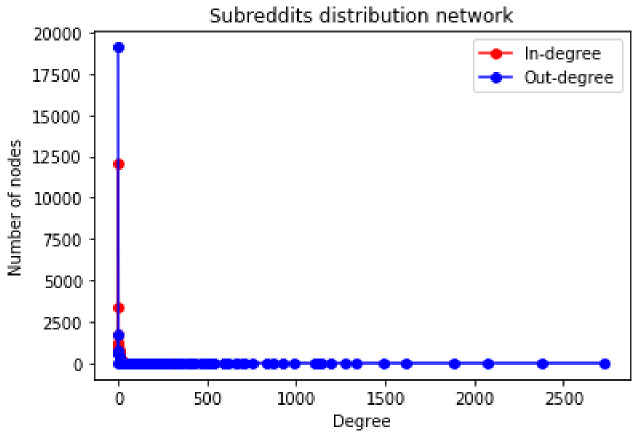

Figure 7: In ad out degree
As shown they are almost identical

with the list not changed, melting in a single row all the information between each pair of nodes that have the same parent and to. All the process could be seen in this .ipyn file "2.1.0)Build- dataframe-Undirected-Graph.ipynb".

Having the new data-frame, we could explore if there is a presence of community in our network. Using different methods present in the CDlib library, we try to understand which community analysis was more suitable for our network. The different algorithm considered were:

- Laiden
- Louviane
- Label Propagation
- Demon
- Infomap

First, we defined the network topology inserting our data and so create a playground to study the diffusive phenomena. For each community discovery algorithm, we obtain an object that implements a concrete instance of the clustering datatype. We calculate clustering parameters[2] if it overlap[3] and the percentage of nodes that the clustering community covers.

Algorithm	Overlap	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

Second, we collect the most interesting clustering evaluation fitness functions. For each community algorithm, was calculated the better internal evaluation through quality scores that are related on different internal community features. Presented below normalized.

Algorithm	size	degree	dens	conductance
Louvian	0.010	0.438	0.269	0.630
Laiden	0.009	0.458	0.138	0.708
Label Prop.	0.147	0.208	1.000	0.543
Demon	1.000	1.000	0.672	1.000
infomap	0.001	0.572	0.672	0.288

Table 5: Fitness Functions Normalized

Observing the conductance between algorithms it seems that Demon and Laiden are algorithms more suitable for our network. However, the nodes covered by Demon are less then the 40% and have overlapping. On contrary Laiden covers all nodes and doesn't take into account the overlap.

Third, given more clustering it could be useful to visualize how a given fitness function distributes over the communities. The internal edge density were compared between all the algorithms, using a violin plot [Figure 9][Figure10].

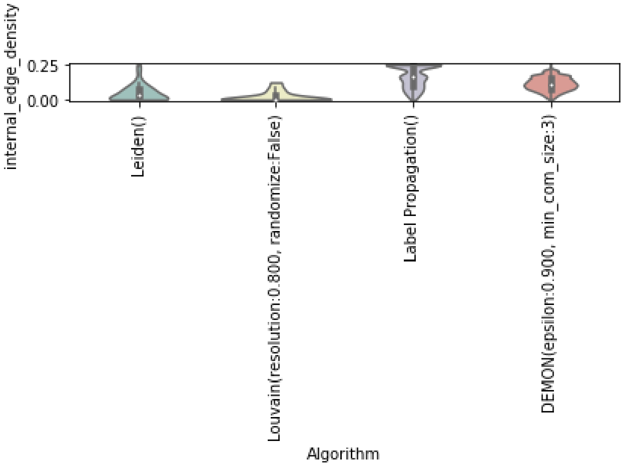

Figure 8: Internal edge density

Form the visualization we could observe that the distribution is better in Laiden community, in fact it has the typical bell form.

We proceed into our analysis, making a qualitative evaluation, analyzing the purity of each community and identifying which have the most homogeneous clusters. This is expressed as modularity which measures the strength of division of a network into modules. Networks with high modularity have dense connections between the nodes within modules but

[2] that vary according to the algorithm used

[3] A clustering is said to be overlapping if any generic node can be assigned to more than one community.

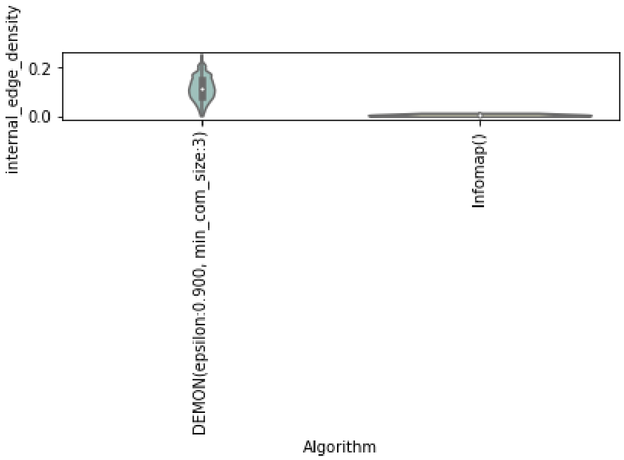

Figure 9: Internal edge density

sparse connections between nodes in different modules. We have used the Erods Reny, Newman Girvan, Z modularity and the density one. Once Normalized we plot them into the graph [Figure 10].

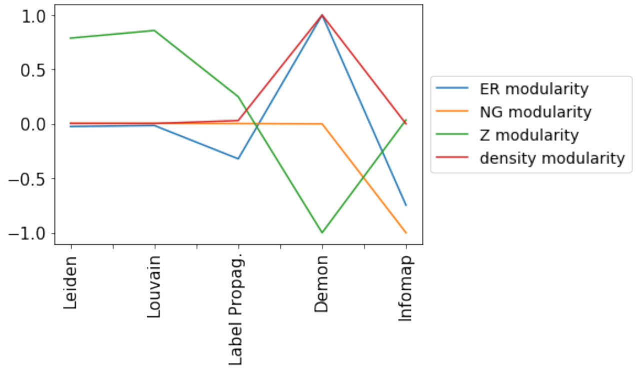

Figure 10: Modularity

As shown, different kind of modularity give us different results, but we could say that in the major cases in our network the different clusters of community are not well separated, in fact most of them are around zero. Above all we observe how the clusters of different algorithms are correlated.[Figure 11]

Their correlations confirm Louvian and Laiden as the most correlated. In conclusion we could say that there isn't a particular algorithm that help us to catch well how the clusters are divided. Despite this, we consider Laiden a quite good algorithm due to his conductance and the distribution that have in the violin plot. So, we could say that, in our network, there are 45 communities that are all quite small (under the 5000 nodes), have internal degree that fluctuate between 5 and 1, a modularity of 0.47 and the fraction of community nodes that belong to a triad is 0.22.

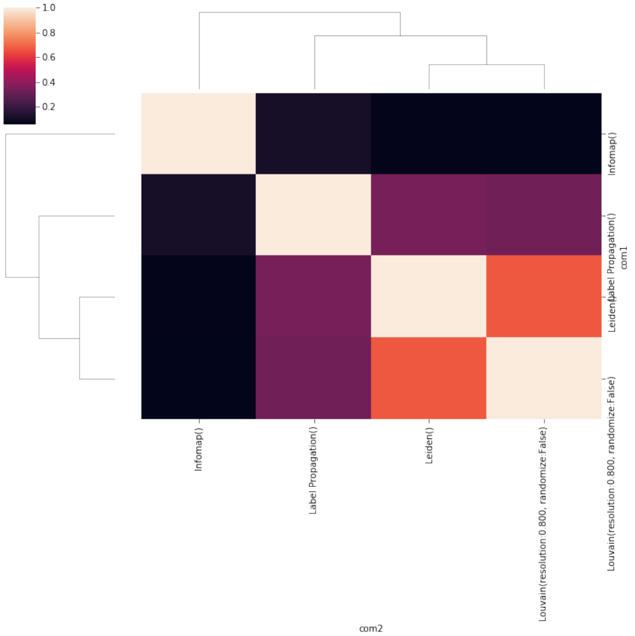

Figure 11: Correlation between all clusters

""" (possibile fare dei 4 grafici in Laiden sulla comunità un unica immagine?)""""

5 TASK 2: SPREADING DIFFUSION

Here we analyze how information, considered as infection, are spreading around our network and compare it with all other synthetic models built before (ER model, BA model, WS model).

We starting our analysis, using the Threshold model introduced by Granovetter[4]. The model works ~~in~~ this way: each node has a threshold; during a generic iteration every node is observed and if the percentage of its infected neighbors is greater than its threshold it becomes infected as well. Using this method we assumed that a meme or a trend as already infected the first percent of our network, for a node is very easy to adopt a 0.2 threshold, although, we iterate the process for 100 times.

"""""(insert img threshold figure ln47)"""""

With this setting the cascade is completed in ER and WS models, instead of our network and in the BA model, the infection stops almost immediately. So, for BA and our network we have tried to start a 10 % of infection ~~supposing that a meme was already very popular.~~ In this case, the simulation shows that the cascade happens for the BA model~~s~~ but

[4] M. Granovetter, "Threshold models of collective behavior," The American Journal of Sociology, vol. 83, no. 6, pp. 1420–1443, 1978

fails to fully happen in our network, ~~it~~ becoming stationary when reach something more then the 80 % of nodes, and thus showing the presence of a cluster with density 1 to 0.2 in our network.

Then we try to explore if there were some differences using the SI model introduced by Kermack[5]. During the course of an epidemic~~s~~, a node change its status from Susceptible (S) to Infected (I). SI assumes that if, during a generic iteration (we set 1500), a susceptible node comes into contact with an infected one, it becomes infected with probability β (in our case set~~tled~~ as 0.001). Once a node becomes infected, it stays infected and at the beginning of the epidemics we assumed the 1% in the network was already infected.

"""(insert img from ln65)"""

The speed of the spreading varies across the models, from faster to slower, we have BA, ER, WS and our network. So due to the fact that our network is the slower we want to observe what could append in the SIS model[6]. The model is same as the one above but a node, instead of remain infected, can switch again to susceptible with probability λ. Here we consider ~~the~~ 1% of nodes already infected, the infection rate equal to the previous one, a recovery rate of 0.005 and $\lambda = \frac{0.01}{0.005} = 2$ (a situation where an outbreak would happen). Iterating all for 600 times.

"""(insert img from ln 84)"""

Here as the one above our network still reach at least 60% of nodes. That's curious, it seems that something in our network doesn't allow to reach the total number of nodes, it could be due to the conformation of our network or lacking of information because we doesn't the complete network of Reddit.

Also we considered a SIS where $\lambda < 1$ so an endemic state is reached $\mu < \beta\langle$k, where$\langle$k$\rangle$is the average degree of our network. We maintain all parameters same as the previous simulation, but changing $\lambda = \frac{0.005}{0.01}$ and a~~dding~~ $0.01 < 0.005 \times 8.63$.

"""(insert img ln105)"""

What we wanted also observe in SIS, a situation where a "disease-free" condition was reached. The condition to achieve this is having $\mu > \beta{<}k$. Setting all parameters with 70% of infected nodes at the beginning, the $\beta = 0.001$, $\gamma =$ 0.05(the recovery rate. We iterate it 100 times.

(insert img ln124)

~~By this graphs we could understand that the infection just could fall down because there aren't anymore much subreddits to infect~~.

It is possible, inside of a subreddit, a post arrived and it is considered interesting by the users of that subreddit, maybe they will re-post it in other, spreading the content, and they will be interested on that posts for a certain number of days. After a while, users will be fed up of that specific topic and introduce other new posts. For observe this, we have used the SIR model[7]. This models follow more or less this concept that a subreddit could be infected, be infected for a certain amount of time and then forget the infection and be healthy again. Here we simulate using this parameters $\lambda > 1$ over the epidemic threshold, infeted at the beginning 1%,$\gamma = 0.005$, $\beta = 0.01$, iterated 1000 times.

"""(insert img ln150)"""

Finally, a simulation where $\lambda < 1$, so below the epidemic threshold, with a infected rate at the beginning of 1%, $\gamma = 0.01$, $\beta = 0.005$ and iterating it 1000 times.

"""(insert img ln174)"""

From the comparison between models, it appears that the spread of the "contagion" in our real network is limited in terms of infected nodes compared to all the other models, furthermore it seems, in ~~the~~ most ~~of~~ cases, that~~,~~ in our network~~,~~ the epidemic is slower in spread in comparison to other models considered.

6 TASK 3: LINK PREDICTION

Unsupervised approach

In this section we discuss the expansion of our network's links between his nodes following the methodology firstly introduced in the article [8] . Using a random model and an unsupervised approach, defining as a set of proximity measures unrelated to the particular network. In particular the predictors used were Random, Common Neighbors, Jaccard, Katz, Graph distance, Page Rank, Simrank. These predictors works differently, 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

[5] W. O. Kermack and A. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 115, no. 772, pp. 700–721, Aug. 1927.

[6] W. O. Kermack and A. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 115, no. 772, pp. 700–721, Aug. 1927

[7] W. O. Kermack and A. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 115, no. 772, pp. 700–721, Aug. 1927

[8] David Liben-Nowell, Jon M. Kleinberg: The link prediction problem for social networks. CIKM 2003

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 classifiers and the accuracy of the prediction done using the edges that gave us the highest score.

Classifier	All data	New data	Weighted data
Random	0.26%	-	-
Common Neighbours	3.70%	5.60%	-
Jaccard	3.31%	4.62%	-
Katz$_{\beta=0.05}$	3.91%	5.57%	-
Katz$_{\beta=0.005}$,	5.27%	**6.95%**	-
Katz$_{\beta=0.0005}$	5.96%	**7.01%**	-
Graph Distance	1.53%	1.81%	0.86%
Page rank$_{\alpha=0.01}$	3.39%	3.91%	4.13%
Page rank$_{\alpha=0.05}$	3.38%	3.94 %	4.17%
Page rank$_{\alpha=0.15}$	3.46%	3.94%	4.22%
Page rank$_{\alpha=0.03}$	3.55%	3.87%	4.19%
Page rank$_{\alpha=0.5}$	3.66%	3.93%	4.22%
Simrank	0.35%	0.55%	0.25%

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.
2. 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.

It is clear that the results are a little bit better, but still too low. The last thing was tring to assign a weight to each link w equal to:

$$w = ln(n_c + n_{up} + 1)$$

where n_c and n_{up} are the numbers of comments and upvotes respectively, 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 improved only by 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. To concluding this part, we obtain a low accuracy due to the unpredictability of the network were the information and content are constantly added and forgotten fast in internet.

Supervised approach

Due to the poor performance of the unsupervised approach we tried to use a supervised one. As features we used the following topological information:

- Jaccard index
- Number of common neighbors
- Adamic Adar
- Preferential attachment

To create the training set we considered all the edges in the training graph (labelling them as 1) and 1.000.000 edges that are not in the core graph (labelling them as 0).

To create the test set we used 1000 edges that are not yet in the training graph, but there will be in the test graph (labelling them as 1) and 200.000 edges that are not in the core graph (labelling them as 0), making sure that we did not use edges already present in the training set.

We created the test set this way, because we wanted to maintain the ratio between class 1 and 0 we would obtain sampling random edges that are not in the training set. As classifiers we used:

- Naive Bayes
- Decision Tree
- Random Forest
- K-nearest-neighbor
- Support vector machine

Before starting the classification we normalized the training and test matrices and applied PCA, to obtain orthonormal ones. For each classifier we computed the precision score,since we are interested on the accuracy of the link prediction, on the test and training sets, obtaining the results shown in table 7:

Even though this is a simplistic approach, we increased or performance by a lot, achieving a precision score of almost 27%.

Classifier	Training	Test
Naive Bayes	25.49%	4.54%
Decision Tree	45.26 %	4.88%
Random Forest	57.41%	8.38%
K-nearest-neighbor	79.14%	**26.96%**
Support vector machine	65.95%	17.59%

Table 7: Precision of the various classifiers

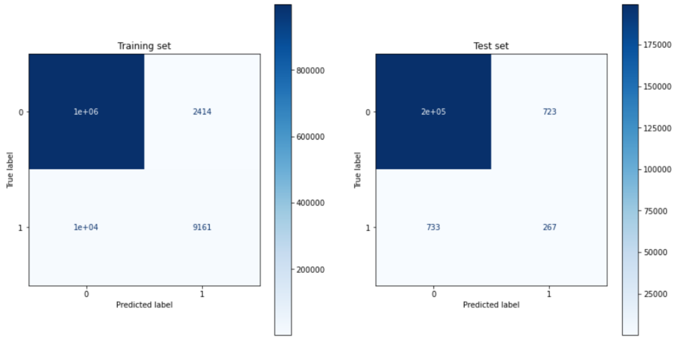

Figure 12: Confusion matrix of the k-nearest-neighbor classifier

7 TASK 4: OPEN QUESTIONS

Each day, on Reddit, people everywhere in the globe, participate and create new subreddits to discuss, share information,content and topic common interests 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 interactions attracts 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 contens, what is the con typology and are their communities 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, their impact on reality and simulate them in a SIS model.

Subreddit characteristics and activities

Yo 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

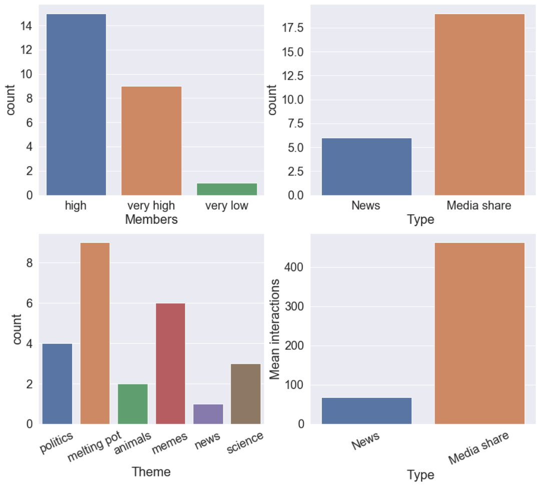

Figure 13: Top 25 subreddits that create more original content

  (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)
  (b) Animals/nature: (example r/awww)
  (c) Science: (example r/technology)
  (d) Memes: (example r/funny)
  (e) News: News subreddits that are not focused on a single topic (example r/news)
  (f) 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 analyzed the distribution of the various class types of the twenty-five subreddits that create most original posts. It emerges that excepting one all the other most active subreddit have an 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.

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 9.25 up-votes on average, if we do 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 9.59 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, to be sure, we also observe his most popular posts crossposted. As we can see from figure 13 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 egnews subreddits generates a lower amount of interactions in respect to the media share subreddits, this is respected in r/ForUnitedStates too.

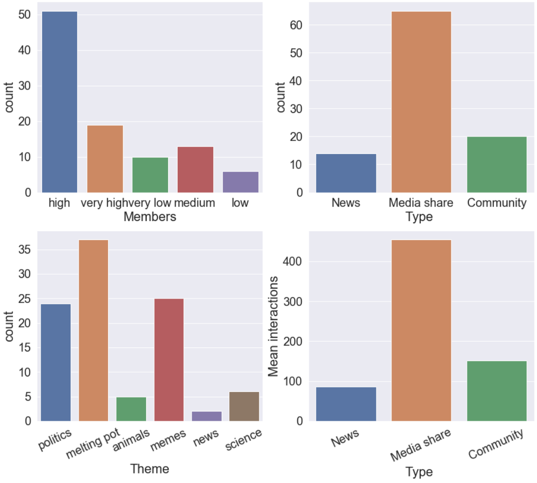

Figure 14: Top 100 subreddits that create less original content

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 reposter 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.

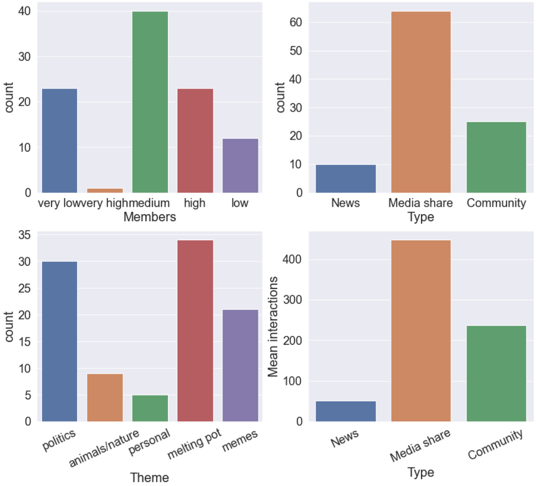

Figure 15: Top 100 subreddits that create more original content

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.

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.

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.

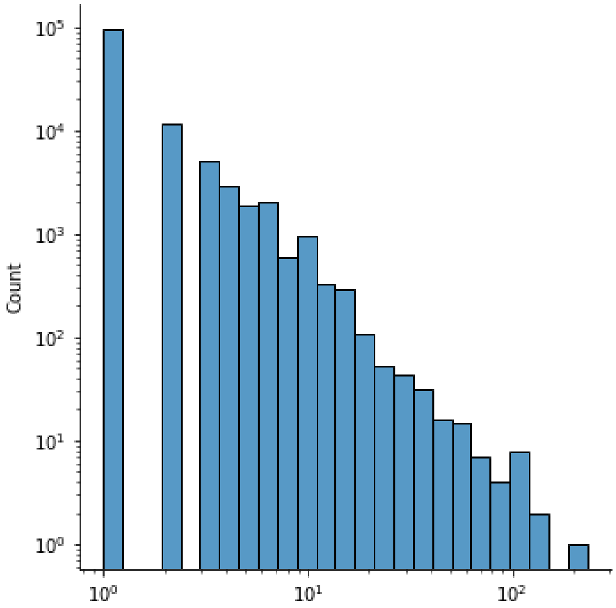

Figure 16: Bilogarithic distribution of the crossposts

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.

Old posts revival

During our data collection we notice some hot posts dated before 2021. That capture our curiosity because we take for granted that the hotted posts will be only ~~the~~ recent one due to the constant bombing of contents that is typical in social media network communities. For this reason we decided to analyze better this kind of posts. What we have done, was picked posts and filtrated in order to select only the ones scraped from the "hot" list of a subreddit. Opposed to all the other cross-posted posts collected after a "hot" crosspost was found. We ordered them by the data of posting and took all posts published before ~~the~~ 2021, using R. After that, with a python script, the members of the subreddits~~,~~ where those posts had been submitted to, were collected with a script [9]. Therefore, $subs_{pre}2021$ contains the number of members of a subreddit with "hot" posts dated older than 2021, 4 months prior to the network scraping. Using that information we plot a distribution. It was immediately clear that the most part of subreditts had a quite modest number of members, eliminated the 132 outliers, this is the distribution:

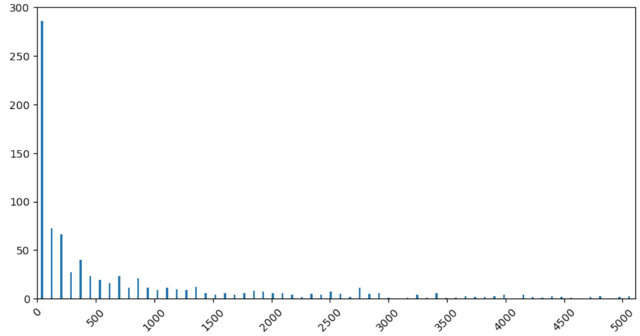

Figure 17: Distribution subreddits/members

Having very old cross-posts among the "hot" ones appears to be a sign of a not very active subreddit. In fact, our code, once find a subreddit, scroll all posts inside it and, if the subreddit is not very active, ~~so~~ it goes back in time until reach the "hot" posts. Randomly, checking the content of our posts, we realised that some of them were pinned[10] and included in the data set. With a python script, we retrieved all the pinned posts and for each subreddit, counted the number of posts not pinned and plotted a distribution.

[9] the number of members of the selected subreddits was collected approximately a month and a half after the initial data scraping. Four subreddits from the partial dataset had been banned in the meanwhile. They have been ignored for the data collection.

[10] A pinned post is a social media post saved to the top of a page or profile on Facebook, Twitter, Reddit and so on. Pinning a post is a great way to feature an important announcement or highlight some of your best content.

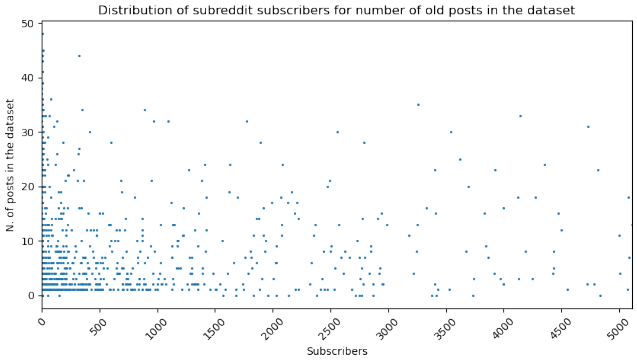

Figure 18: Distribution subreddits posts

As already explain above, the imagine well confirmed that small subreddits have less active users, in fact it is possible to observe an high distribution of old posts in them.

Game Stop Case, spreading in our network

On January 2021 the price for Gamestop stocks reached an all time high of 440$. This happened after the members of the reddit community r/wallstreetbets, started buying a large quantity of the title, causing a chain reaction that led to the high price. Once the price rose, the news starded spreading through the social network and new memes and discussions arose from all over the platform. What we have in mind is try to model the spreading of those posts using a simple SIS model. To retrieve the data we used pushift, a reddit archive. We found all posts created between 1st January 2021 and 1st March 2021 containing the words GME or gamestop in the title, saving those posts into an infection database. Once a subreddit contains a post about GME, it is considered infected. After 15 days from the last post we considered it susceptible. What was curious is that the majority of our graph was not infected. This may be due to the fact that there is a very low chance of infection for subreddits that are not interested in finance or memes (ex. r/ferrari). We try to consider only subreddits that are in the same cluster of r/wallstreetbets, for observing if something change, but the infection wasn’t so much improved [Figure 14].

Due to the fact that the infection wasn’t spread so much, we have decided to analyze the relationship between the number of infected subreddit and the logarithm of the stock price.

Clearly, we could say the infection is not bounded with the stock market price, in fact people talk about their price just when this have reach an important quotation in the stock market. Nevertheless the interests die when the quotation in the stock market is higher then in the period where everybody are interested in that topic.

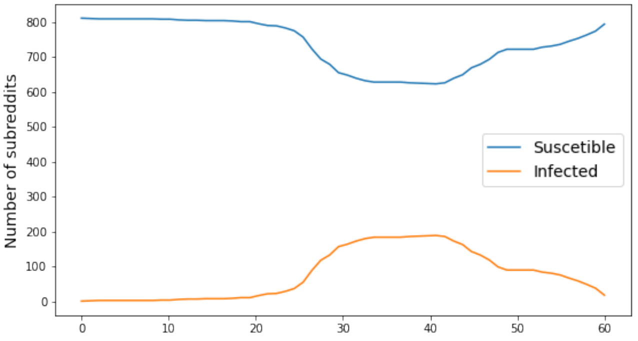

Figure 19: The Game Stop infection in r/wallstreetbet

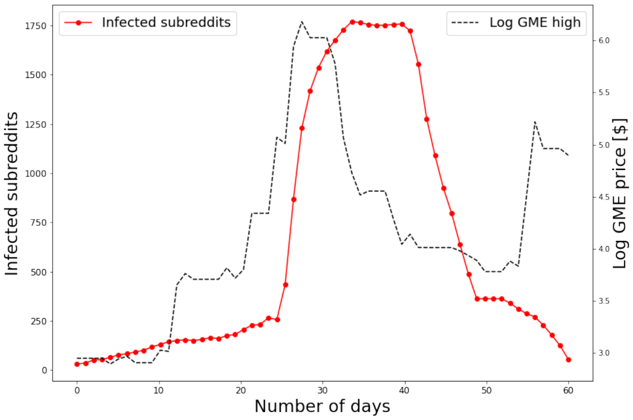

Figure 20: number of infected subreddit and the logarithm of the stock price

Using the SIS model, we try to fit the data, following this function:

$$i(t) = (1 - \frac{\mu}{\beta})\frac{Ce^{(\beta-\mu)t}}{1 + Ce^{(\beta-\mu)t}}$$

and consider that it could be present an error. The idea for construct an error considers that it is easier miss an infected subreddit while the infection is in his early stage. So, we set some criteria:

- 5% of the total infected for days with more than 250 infected.
- 20% for days with less than 250 infected.

Other things to consider are the parameters β, μ that have to be constant with time. But the first model consider wasn’t a good fit. In view of have the best fit possible, we considers that during the epidemics something may have happened that chanced the spreading parameters, mainly the drastic price increase of the GME stock. For this reason the time domain was in 3 different regions:

(1) A linear phase from 01/01 to 25/01
(2) An exponential increase from 25/01 to 09/02

	First Phase	Second Phase	Third Phase
β_t	0.3 ± 0.1	0.48 ± 0.02	0.04 ± 0.02
μ_t	0.3 ± 0.1	0.44 ± 0.02	0.05 ± 0.02
R_t	1.0 ± 0.5	1.08 ± 0.08	0.9 ± 0.6
χ^2/ν	1.05	0.94	11.92

Table 8: Caption

(3) An exponential decrease from 10/02 on wards

For each of those phases were computed β_t, μ_t, R_t.

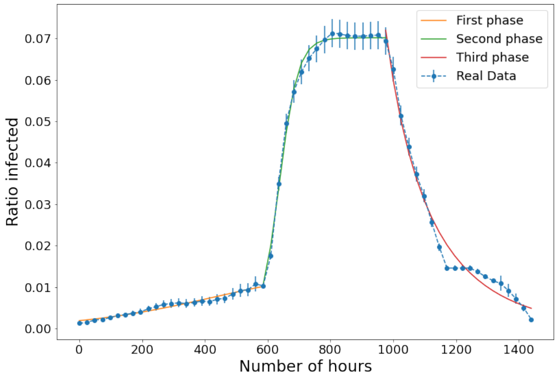

Figure 21: fitted data

The first two model fit the data quite well, in fact, we have a value of $\frac{\chi^2}{\nu} \approx 1$, while the last model does not yield good results.

One of the main assumption on epidemic spreading on social networks is that a pathogen (in our case a meme) can be spread only to its neighbors. We try to to test if this assumption is true and to do so, we take a look closer to the number of weakly connected components in the infected subgraph. Considering that: once a subreddit is infected it stays infected, so what happen to the number of weakly connected components? Should remain the same or decrease in time?

Surprisingly, the opposite is true. This mean new infections do not stem from the infected node neighbors, but pop up randomly. Why this happen? Probably due to 3 different reasons. First, our network is only a static snapshot of a time-varying and more complicated network. Second, the number of subreddits that are infected is much higher than what we found. Third, our network is not complete, we are missing some key subreddits such as the presence of the subreddit r/all. In fact, differently for the majority of social networks (in which a user can see only see posts of friends or pages he follows), in reddit there is a special page called r/all, that contains all the most popular posts of the day. This means each reddit user can see news and memes about game stop, thus meaning each subreddit can be infected even if it has no infected neighbors.

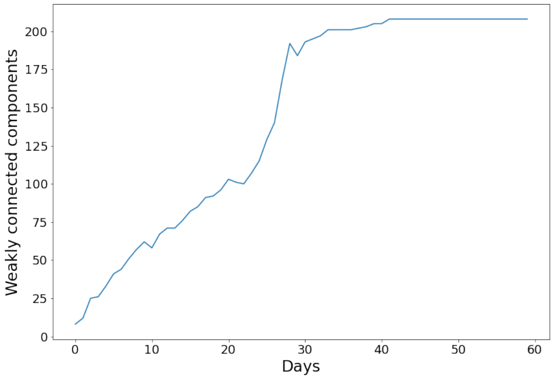

Figure 22: infection throughout the weakly connected components

For complete the last part of our exploration, we ran two models of SIS simulation. The first one considered have an infection rate of 0.359, and a recovery rate of 0.333.

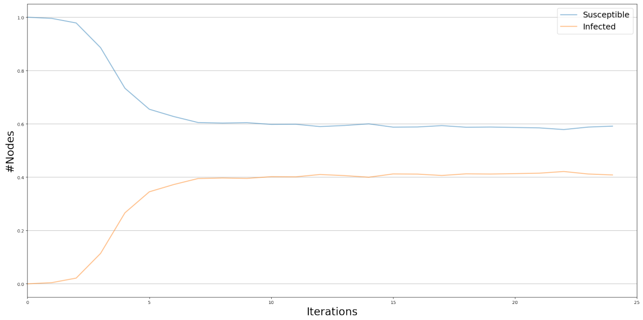

Figure 23: First Model

That's wasn't a satisfied infection and the data not fitted well. In fact, the infection reach at least ~~the~~ 40% of the nodes. To reach ~~a~~ better results we try to divide our model in different phases of infection. In the first phases we consider the ~~same~~ infection rate of 0.298 and the 0.293 the recovery rate. Also here wallstreetbets subreddit is the starting point for the epidemic.

Figure 24: Second Model, First phase

For the second phase we have used an infection rate at 0.478 and a recovery rate at 0.444. This are higher due to the fact that in a second phase it is more easy be infected then the first one. ~~What's more~~ we ran models that start~~s~~ with different percentage of infected subreddits. The differences were the blue one start at 1% of infected subreddits, the yellow represent the beginning with 40% of infect, the green is a 70% of infected, red consider how many neighbours of a node were already infected and the purple take into account the the nodes that were infected in the first phase.

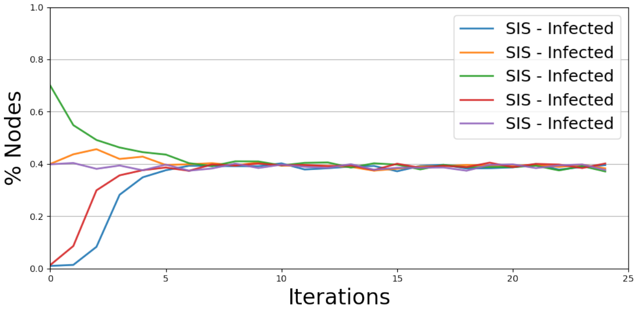

Figure 25: Second Model, Second phase

At this point, we started to think that the % of infected nodes and the topography of our graph, influence heavily the way in which the infection spreading plays out.

In the third phase the recovery rate and infection rate were respectively 0.048 and 0.043. This time the different percentage were chosen in one case because we observe that mathematically that it as to be a 7% of infected subreddits. The other two used were the 40% and the 70% of infected subreddits.

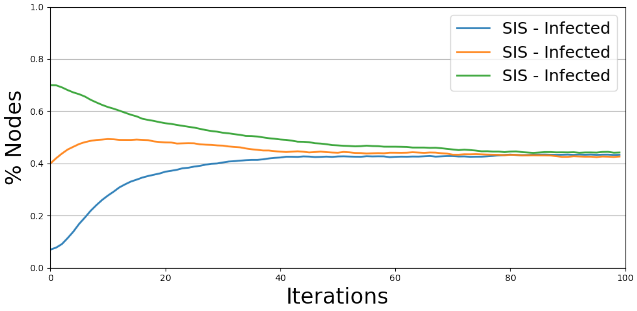

Figure 26: Second Model, Third phase

Also here, no matter what, the infection continue to converge towards a certain equilibrum set around 40%. Similarly we have ran other simulation giving some wadges or consider the interactions between subreddits but the reaseach do not give to us satisfied prof. In conclusion it seems that the simulations fail because sure we do not have a complete graph, maybe for our network the SIS analysis could be not the suitable one. In addiction the not possibility that in our network the infection reach all the nodes could be caused by a sort of background noise that modify and change the network spreading.

8 DISCUSSION

REFERENCES