How trends spread through reddit

Bettini Chiara
c.bettini3@studenti.unipi.it
Univeristy of Pisa, Student ID: 518134

Lestini Cinzia
21980920@studenti.unipi.it
Univeristy of Pisa, Student ID: 219809

Manassero Chiara
c.manassero@studenti.unipi.it
University of Pisa, Student ID: 547922

Spina Paolo
p.spina4@studenti.unipi.it
University of Pisa, Student ID: 568773

ABSTRACT

In this paper we analyzed one of the most unique social network in the world: Reddit, focusing on the relationship between subreddits. In order to do so, we created a network, in which the nodes are the subreddits and the links are shared posts between them, also known as crossposts. We tried to model how information spread between subreddits and surprisingly we found out that viral posts (post that generates a large number of interactions such as upvotes and comments) do not span across multiple subreddits. This prompted us to shift our attention from single posts to a larger picture. We tried to model the diffusion of news and memes about how r/wallstreetbets caused a chain reaction back in January 2021, that caused the price of GME stocks to reach an incredible value of 350$. We found out that information do not spread through neighbors, like other social network, but in a seemingly random manner. This peculiarity allowed us to use a standard SIS model with very good results.
[1]

KEYWORDS

Social Network Analysis, Reddit

ACM Reference Format:
Bettini Chiara, Lestini Cinzia, Manassero Chiara, and Spina Paolo. 2021. How trends spread through reddit. In *Social Network Analysis '21.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[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 INTRODUCTION

In this paper we explore and comment all the 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 cross-posted also after one or two year later? How is it 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 why we wanted to use 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 and data source environments in which was possible to obtain data. After checking the feasibility or not of different topic, excluding some of them due to: lack of number of nodes, impossibility to reach information, we agreed to use Reddit as our environment to extract data and try to observe and study his communities 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 formed by people to discuss common shared interest and they also could include 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 on specific one such as r/hockey. Secondly, few months ago, into a financial subreddit people found an arbitrage opportunity in the stock market and bought GameStop

stocks until the price reached 350$. This action lead to a chain reaction that never seen before in the stock market. 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), and, thanks to python's praw library, we were able to use their interface easily. Lastly, python's library praw allow us to reorganize the extraction, collection and clean all reddit's data. What we have in mind was observe how subreddits could interact one to each other. The presence of a cross-post[2] could be consider an interaction between two subreddit. 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
- subreddits as nodes
- crossposts as links

Crawling Methodology and Assumptions. The major code created 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 these 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 (subreddit in which the post firstly appears)

These 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, we have iterated the process until we reach the third level. At the end, we built a code to delete redundant elements and possible parallel edges, obtaining a cleaned file.

3 NETWORK CHARACTERIZATION

To characterize our network we used the NetworkX library. Our network 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. Furthermore, we were able to calculate all these following characteristics of our network:

As we can notice the network is a medium size one with 24819 nodes. On average nodes have a degree of 8.63. It is important to underline that the average degree is divided between in-degree and out-degree because we are observing a directed network. 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. 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, so we have a sparse network. Our major hubs 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 less original content are: "GoodRisingTweets", "LateStageCapitalism", "aww".

	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

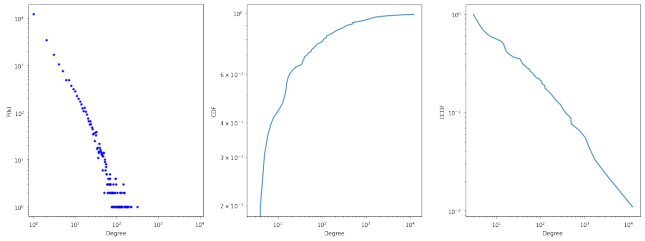

Figure 1: Degree distribution
Observing the first graph on right it emerges that we have a network with some few nodes heavily connected and lot nodes that posses just few connections.

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.

[2]For cross-post we mean the process of posting the same Reddit post in several different subreddits

Comparison with ER and Configuration model

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 $L = p * [N(N-1)]$, 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

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

.

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.

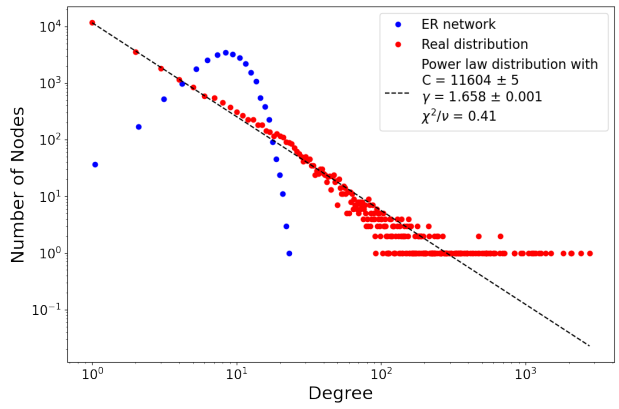

Figure 2: Degree distribution Subreddit network and ER

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.

According to the figure, it seem to follow a power-law distribution. 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. Under a certain value we divided data in a range of values and for each rage we picked the average one. The results confirms our idea. From the comparison it emerges that ER model is in 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 was built giving to each node a particular in and out degree. But it fails in the representation of the average path length. In conclusion we could say that ER model is situated in a super-critical regime, have a big one giant component, that present loops and self loops. Nontheless the Configuration model is similar to our network due to the degree distribution. That wasn't enough so we tried to explore more with Watt-Strogatz Model.

	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

Comparison with the Watt-Strogatz Model

Watts-Strogatz[3] 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 8.63. Then we made two models with the same number of nodes as our network (24819), but setting k=8 for the first and k=10[4] for the second. Doing this we obtain different number of links for each value: k=8 gave us 99276 links and k=10 had 124095. As predictable, no one of the two limits generated the same number of links as our real network. For both these k, we have to create and calculate their major feature using a Watts-Strogatz model with different p: a Regular Lattice, a random graph and graph under the small world regime. we calculated the major characteristics setting for each the upper and lower bounds.

On one hand, Regular Lattice models have an average clustering coefficient significantly higher than our network. On the other hand, Random models have an average clustering coefficient lower than ours. Small World Regime is near to

[3] Duncan J. Watts and Steven H. Strogatz, Collective dynamics of small-world networks, Nature, 393, pp. 440–442, 1998

[4] k=10 because it is not possible to build a synthetic graph with and odd number of edges

Table 3: Table following Watts-Strogatz Model

	Lattice	Random	Small World
lower, upper	k = 8 | k = 10	k = 8 | k = 10	k = 8 | k = 10
p	0	1	0.37
<C>	0.6428 | 0.6667	0.0003 | 0.0004	0.1644 | 0.1695
k	8 | 10	8 | 10	8 | 10
<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

our network in the average clustering and the average shortest path, characteristic that is also similar to our graph and the Random model. Small World Graph is the most similar model to our Subreddit network but the density is different. We could try using a Barbarasi-Albert model to explore deeply our network.

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 Barabarasi-Albert graphs (BA), indirect. We create this 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). It's visible, in Figure 4, 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 distribution3.

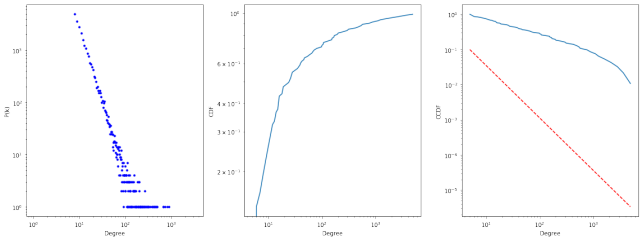

Figure 3: BA model degree distribution
The synthetic BA model has alpha=2.505 and sigma=0.049.

Just the diameter and the degree distribution of BA model are same as our, for the rest the average degree is double and also the density. 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

Model	BA	Subreddit
Configuration	m=8 N=24819	-
N	24819	24819
L	198488	104284
<K>	15.99	8.4036
K_{max}	703	2754
K_{min}	8	1
density	0.0006	0.0003
<C>	0.0046	0.1654
<d>	4.1834	3.7362
distribution	power law	power law

Table 4: Comparison between Barabasi-Albert and Subreddit network

presence of hubs as was already pointed out in the first part of the network characterisation.

4 TASK 1: COMMUNITY DISCOVERY

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 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 for each pair of nodes. What's more, we integrate the modified list 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. Having the new data-frame, we could explore if there are communities in our network. Using different methods in the CDlib library, we try to understand which community analysis was more suitable for our network. The different algorithms considered were:

- Leiden
- Louviane
- Label Propagation
- Demon
- Infomap

First, we defined the network topology inserting our data and 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[5], if it overlap[6] and the percentage of nodes that the clustering community covers.

Algorithm	Overlap	Node cover
Louvian	no	1.0
Leiden	no	1.0
Label Prop.	no	1.0
Demon	yes	0.3857
Infomap	no	1.0

Table 5: Community Discovery Algorithms

Second, we collected the major clustering evaluation fitness functions. For each community algorithm, was calculated the internal evaluation, considering quality scores that are related on different internal community features.

Algorithm	size	degree	dens	conductance
Louvian	0.010	0.438	0.269	0.630
Leiden	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 6: Fitness Functions Normalized

Observing conductance between algorithms it seems that Demon and Leiden are algorithms more suitable for our network. However, nodes covered by Demon are less than 40% and have overlapping. On contrary, Leiden covers all nodes and doesn't take into account the overlap. Third, it could be useful to visualize how a given fitness function distributes over the communities. Internal edge density were compared between all the algorithms, using a violin plot [Figure 4].

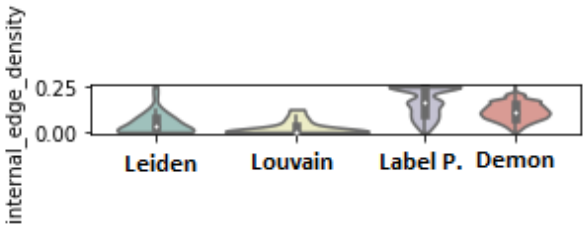

Figure 4: Internal Edge Density
Infomap algorithm was excluded because give us not convinced data as is possible to observe in [Table 6]

[5]that vary according to the algorithm used
[6]A clustering is said to be overlapping if any generic node can be assigned to more than one community.

Form the visualization we could observe that the distribution is better in Leiden community discovering algorithm, in fact it has a bell form.

We proceed into our analysis, making a qualitative evaluation, observing which algorithm have the most homogeneous clusters. Using modularity we measured the strength of division of a network into modules. Networks with high modularity have dense connections between nodes within modules but sparse connections between nodes in different modules. We have used the Erods-Renyi, Newman Girvan, Z modularity and density modularity. Once Normalized we plot them into the graph [Figure 5].

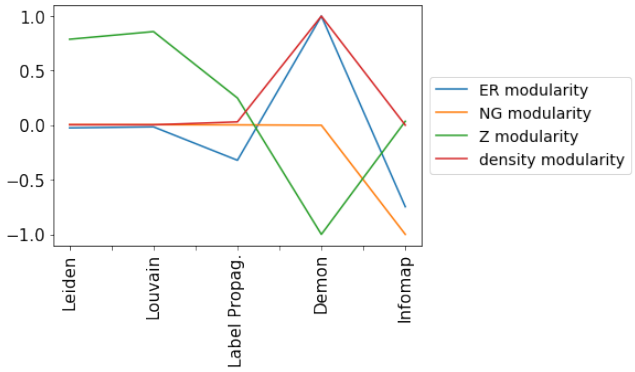

Figure 5: Modularity

As shown, in the major cases, in our network there is not a easy distinction between the density inside communities and outside them. Above all, we observe how the clusters of different algorithms are correlated.[Figure 6]

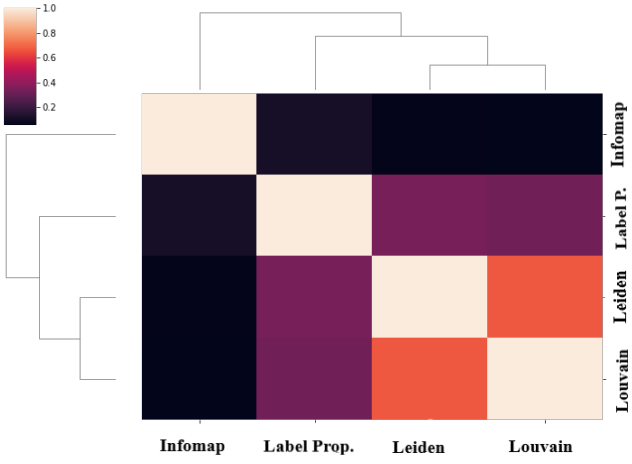

Figure 6: Correlation between algorithms

In conclusion we could say that there isn't a particular algorithm that catch well clusters division. Despite this, we

consider Leiden a quite good algorithm due to his conductance and the distribution that have in the violin plot. According to this, in our network, we observed 45 communities that are all quite small (under the 5000 nodes), have an internal degree between 5 and 1, a modularity of 0.47.

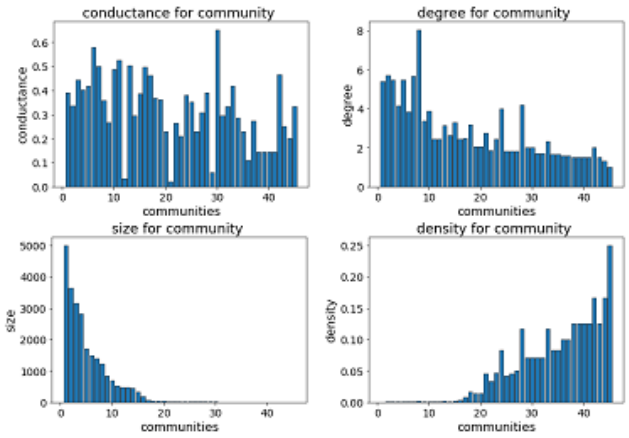

Figure 7: Leiden Communities Discovery

5 TASK 2: SPREADING DIFFUSION

Here, we analyze how information, considered as an infection, are spreading around our network. Compare these results ran on our network with all other synthetic models built before (ER model, BA model, WS model).

We started our analysis, using the Threshold model introduced by Granovetter[7]. The model works 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 a meme or a trend has already infected the first percent of our network. We set a low threshold and iterated for 100 times.[Figure 8] With this setting the cascade was completed in ER and WS models. In our network and in BA model, the infection stops almost immediately. Due to this, BA and our network were ran setting a 10 % of infected at the beginning. In this case, the simulation shows that the cascade happens for the BA models but fails to fully happen in our network. It becoming stationary after reaching the 80 % of nodes. Thus shows the presence of a cluster with density 1-0.2 in our network.

Then we try to explore if there were some differences using the SI model introduced by Kermack[8]. During the course

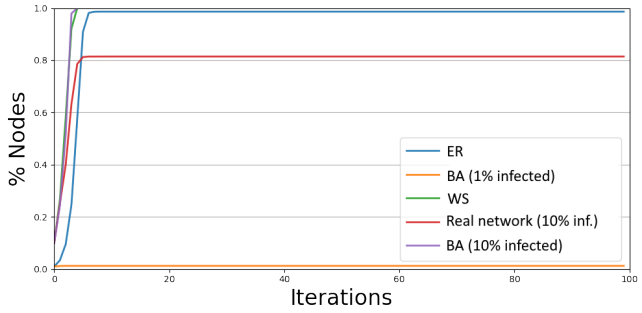

Figure 8: Threshold Model

of an epidemics, a node changes 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 as 0.001). Once a node becomes infected, it stays infected and at the beginning of the epidemics we assumed the 1% of the network was already infected.[Figure 9] The speed of the spreading varies across the models, from

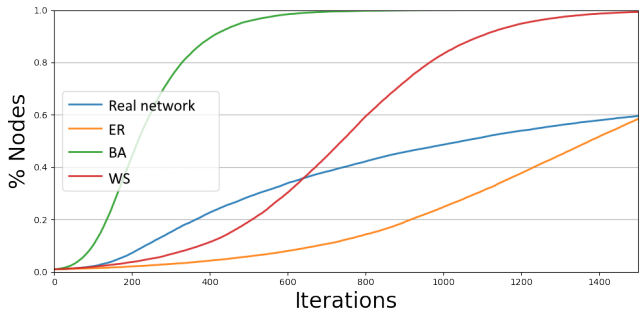

Figure 9: SI infection

faster to slower, we have BA, ER, WS and our network. So our network is the slower. We wanted to observe if using the SIS model[9] something changes. The model is same as the one above but a node, instead of remaining infected, can switch again to susceptible with probability μ. Considering 1% of nodes already infected, the infection rate equal to the previous one, a recovery rate of 0.005. Therefore the basic reproductive number $\lambda = \frac{0.01}{0.005} = 2$ (a situation where an outbreak would happen). We ran the simulation for 600 iterations. Here, as the one above our network still reach at least 60% of nodes.[Figure 10] We considered a SIS where an endemic

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

[8] 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.

[9] 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

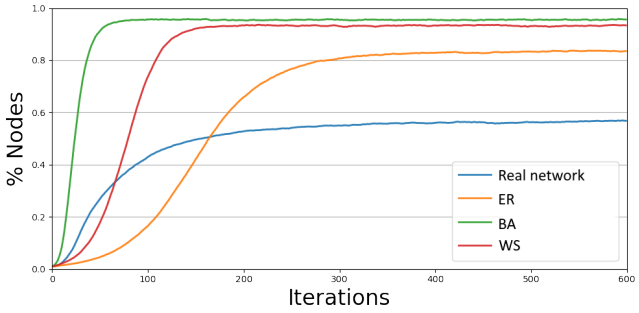

Figure 10: SIS infection λ above one

state is reached, i.e where $\mu < \beta$ <k>, where$\langle k \rangle$is the average degree of our network. Parameters this time were:μ=0.01, β=0.005. Therefore $\lambda = \frac{0.005}{0.01}$ and $0.01 < 0.005 \times * 8.63$.[Figure 11] Finally, we ran a SIS simulation where a "disease-free"

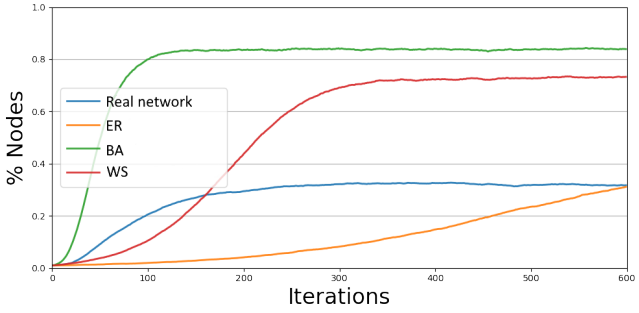

Figure 11: SIS endemic infection

condition was reached. The condition to achieve this is having $\mu > \beta < k >$. In this case parameters where: 70% of infected nodes at the beginning, the $\beta = 0.001$, $\mu = 0.05$. We iterate it 100 times.[Figure 12] It is possible, inside of a

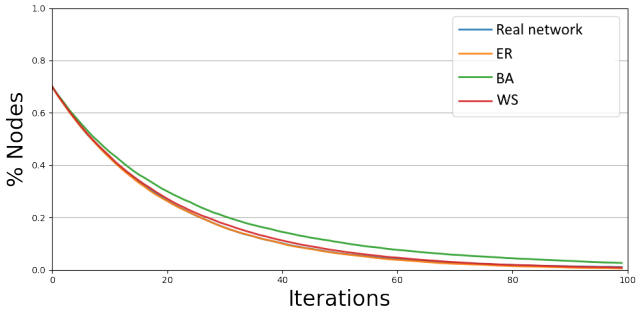

Figure 12: SIS disease free infection

subreddit, a content arrived and it is considered interesting by the users of that subreddit. Maybe they will re-post it in other, spreading the content. After a while, users will be fed up of that specific topic and forget about it. For observe this, we have used the SIR model[10]. SIR assumes that if, during a generic iteration, a susceptible node comes into contact with an infected one, it becomes infected with probability β, than it can be switch to removed with probability μ. First, we run a simulation using these parameters: $\lambda > 1$ over the epidemic threshold, $\mu = 0.005$, $\beta = 0.01$, 1% of the network already infected, iterating 1000 times[Figure 11]. Finally, we

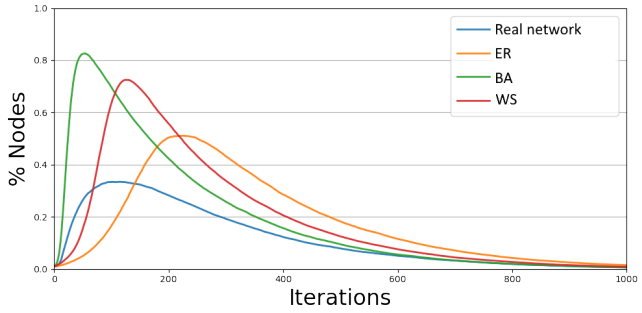

Figure 13: SIR infection λ above the epidemic threshold

ran a simulation where $\lambda < 1$, so below the epidemic threshold, with a infected rate at the beginning of 1%, $\mu = 0.01$, $\beta = 0.005$ and iterating it 1000 times.[Figure 12] From the

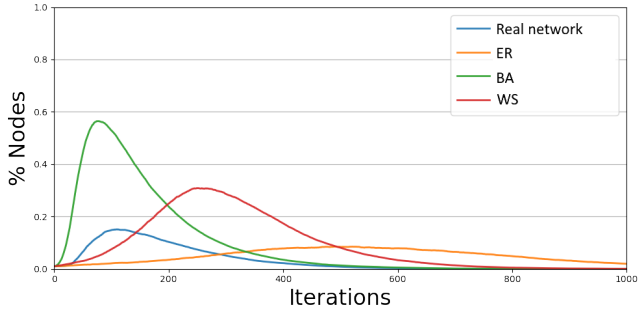

Figure 14: SIR infection with λ below the epidemic threshold

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 most cases, that in our network, the epidemic is slower to spread compared to other models considered. It seems there is something that stops the infection or it could be due to the conformation of our network and the lacking of information because we do not have the complete network of Reddit.

[10]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 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 [11] . 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 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 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.

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%. 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.

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

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

Table 7: Link prediction accuracy

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 trying 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 up-votes respectively, following the idea that a cross-post 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 following the procedure used in this article [12]. As features we used the following topological information computed on the training graph:

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

[12]Link Prediction using Supervised Learning by Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki

To create the classification dataset we sampled 2.000 edges in the test graph (labelling them as 1) and 400.000 edges that are not in the core graph (labelling them as 0).
We created the classification dataset 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

We used a standard 5-fold cross validation and set up a classification pipeline composed by a scaler to normalize the input data, PCA to obtain an orthonormal feature matrix and a classifier. For each classifier we computed the average precision and recall scores on the test set.The results are shown in table 8:

Classifier	**Precision** [%]	**Recall** [%]
Naive Bayes	5.5 ± 0.4	26 ± 2
Decision Tree	3.1 ± 0.1	30 ± 1
Random Forest	3.8 ± 0.4	22 ± 3
K-nearest-neighbor	**61 ± 4**	21 ± 3
Support vector machine	8 ± 5	0.2 ± 0.1

Table 8: Precision and recall of the various classifiers

One of the main challenge of this dataset was the highly imbalanced classes, but we managed to obtain a precision of 61% using a knn classifier, a result much better than the one obtained using an unsupervised approach. To further improve this result, it is possible to do link prediction considering only nodes of the same community, to reduce the imbalance betwenn class 0 and class 1, due to the fact that communities have higher density.

7 TASK 4: OPEN QUESTIONS

Each day, on Reddit, people everywhere in the globe, participate and create new subreddits to discuss, share information, content common interest 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 our attentions and now, that we have a better picture of the network built, we asked ourselves:

- Which are the subreddits more active in publishing and share original content? what is the content typology? The subreddits that are more likely to share less original content which characteristics have?
- 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 the viral post GME and Game Stop, could spread in our network and simulate them using the SIS model.

Subreddit characteristics and activities

To observing subreddits characteristics and activities, we choose in a first moment, to explore the first one-hundred subreddits that post most original content. Before starting we classified them according to their:

- Number of members

very high	> 10.000.000
high	> 1.000.000
medium	100.000
low	> 10.000
very low	< 10.000

- Type
  (1) News: Subreddits focused on sharing news (example r/environment)
  (2) Media share: Subreddits focused on sharing media like images, videos, gifs etc (example r/memes)
  (3) Community: Subreddits composed of people sharing a common interest (example r/gaming)
- Theme
  (1) Politics: (example r/Europe)
  (2) Animals/nature: (example r/awww)
  (3) Science: (example r/technology)
  (4) Memes: (example r/funny)
  (5) News: News subreddits that are not focused on a single topic (example r/news)
  (6) Melting pot: There is not a main theme, but rather a mixture of all the themes I listed above (example r/BeAmazed)

Once the data-frame was built with this categorization, we analyzed the distribution of the various class types of the first 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. We have taken a closer look on this subreddit may be caused by a SPAM or by dedicated users.It certainly was cross-posted in other subreddits 1929 times and it had created 405 original post. 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

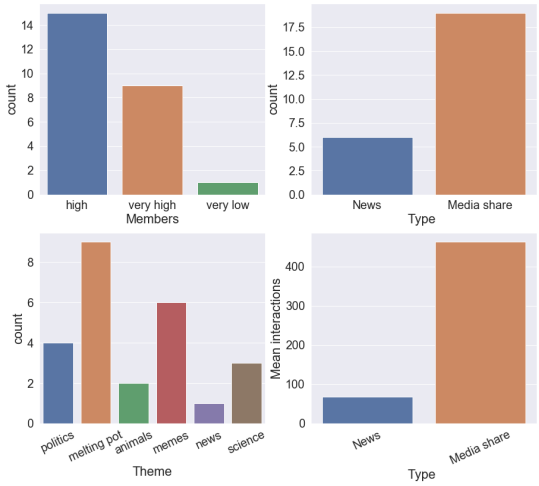

Figure 15: Top 25 subreddits that create more original content

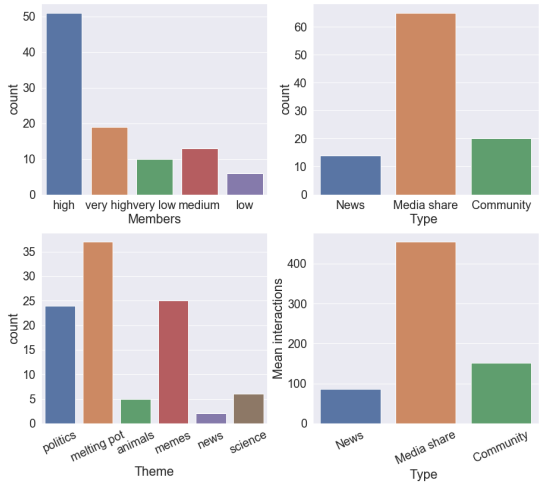

Figure 16: Top 100 subreddits that create more original content

raise up to 3.53 and up-votes did not change practically because it is 9.59 the average. On first impact, do to his small amount of comments and interactions seems to be a SPAM subreddit but, to be sure, we also observe his most popular posts crossposted. As we can see from figure 15, comparing r/ForUnitedStates with all the other 25^{th} subreddits in average interactions and type, we could say this subreddit is not a SPAM because, if we filtered the results by subreddit type, we can see that all the news subreddits generate a lower amount of interactions in respect to the media share subreddits, this is respected in r/ForUnitedStates too.

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 re-poster we could say the same things about the subreddit that on average share more crossposts then original ones?

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

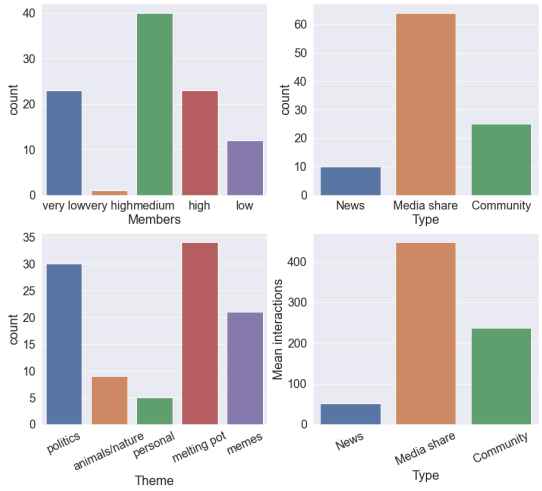

Figure 17: Top 100 subreddits that create less original content

subreddits but emerges personal one. Those are private subreddits created and operated by a single redditor, in which all posts are cross-posts 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. There is only one subreddit with number of member over 10M, 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 contents are created.

As shown before, we could say the major part of original posts 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? In order to figure how this could happen, we were thinking that probably the more up-votes a post has, higher is the probability of this post to be seen and cross-posted. Surprisingly the number of subreddits its very low. In fact,

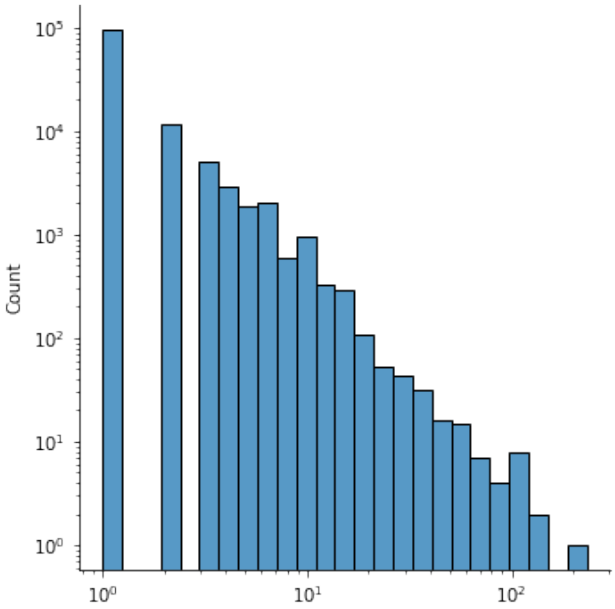

Figure 18: Bilogarithic distribution of the crossposts

the majority of the posts were cross-posted in only one other subreddit. According to our observations, it is 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. Not to mention, the subreddit r/all, in where are published all the viral content and every reddit users could see them, could alter the process of cross-posting because a content could be cross-posted by a subreddit directly form the subreddit r/all.

Old posts revival

During our data collection we noticed some hot posts dated before 2021. That captured our curiosity because we took for granted that the hottest posts will be only recent ones 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. First, we filtrated the posts 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" cross-post was found. We ordered them by the date of posting and took all posts published before the 2021, using R. After that, with a python script, the number of members of the subreddits, where those posts had been submitted to, were collected[13]. Therefore, we obtained a data-set with the number of members of a subreddit with "hot" posts dated older than 2021[14]. Using that information we plotted a distribution. It was immediately clear that most of these subreddits had a quite modest number of members. Eliminating the 132 outliers, this is the distribution:

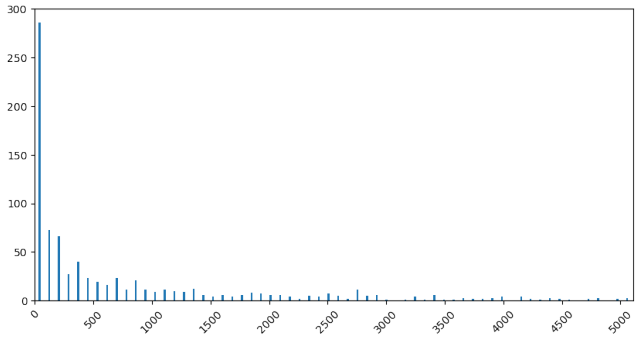

Figure 19: Distribution subreddit/members

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 it finds a subreddit, scrolls all posts inside it looking for cross-post and, if the subreddit is not very active, it goes back in time until reaches the "hottest" posts. Randomly, checking the content of our posts, we realised that some of them were pinned[15] and included in the data set. With

[13] 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.

[14] collected 4 months after the network scraping

[15] 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.

a python script, we retrieved all the pinned posts and for each subreddit, counted the number of posts not pinned and plotted a distribution.

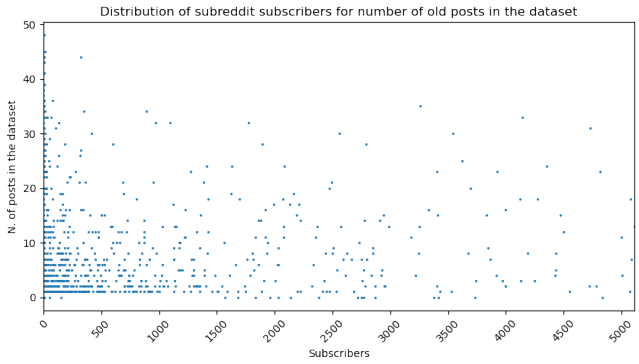

Figure 20: Distribution of subreddit subscribers for number of old posts

As already explained above, the imagine well confirmed that small subreddits, in our data-set, are less active, in fact, it is possible to observe an high distribution of "hot" 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 350$. 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

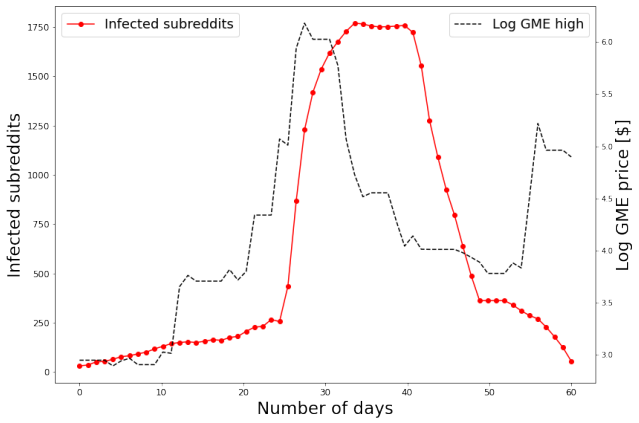

Figure 21: number of infected subreddit and the logarithm of the stock price

	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 9: Caption

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.

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

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

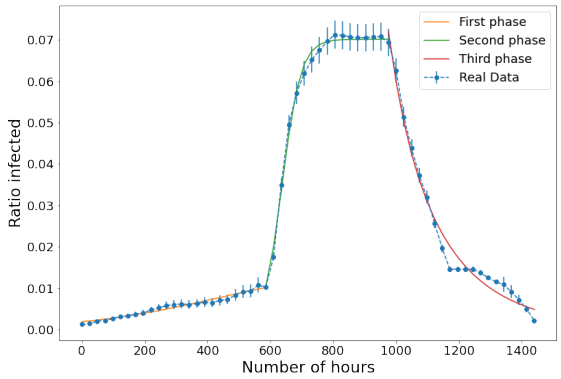

Figure 22: fitted data

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.

To explore the spreading infection of the Game Stop content in our network we ran two SIS simulation with the parameters found observing the real data. The first one considered has an infection rate of 0.359, and a recovery rate of 0.333.

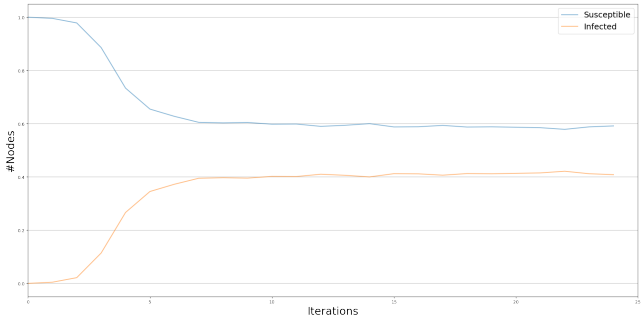

Figure 23: First Model

The model didn't fit well the data and the simulation didn't perform well either. In fact, the infection reaches at least 40% of the nodes. We then tried the parameters of the second model, dividing it in three phases of infection. For the first phase, we ran a simulation with an infection rate of 0.298 and a recovery rate of 0.293 and setting r/wallstreetbets a starting point for the infection.

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 easier be infected than in the first one. We ran simulations with 1%, 40% and 70% of subreddits alredy infected, one with a nodes that had

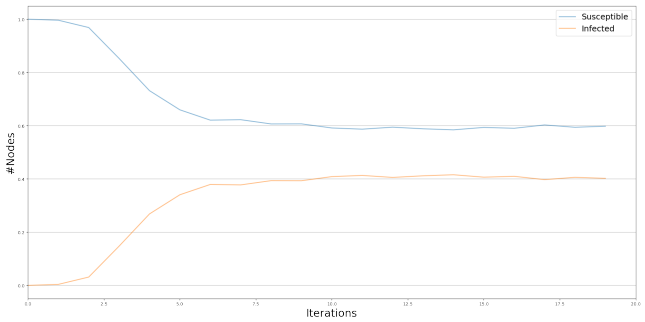

Figure 24: Second Model, First phase

neighbours already infected and one with the nodes that were infected at the end of the simulation in the first phase.

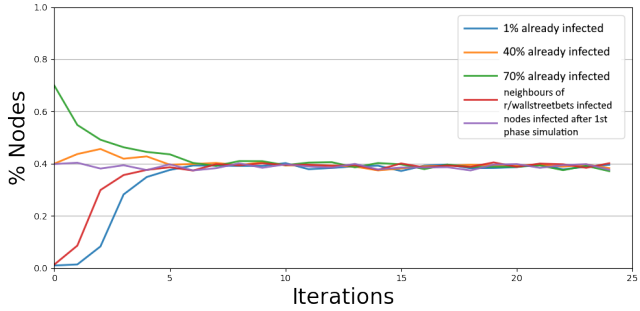

Figure 25: Second Model, Second phase

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, we ran a test with 7% of the subreddits infected to reflect the real data we have previously analyzed. Another simulation was ran with 40% of nodes infected to reflected the simulation of the second phase and one with 70% of nodes infected.

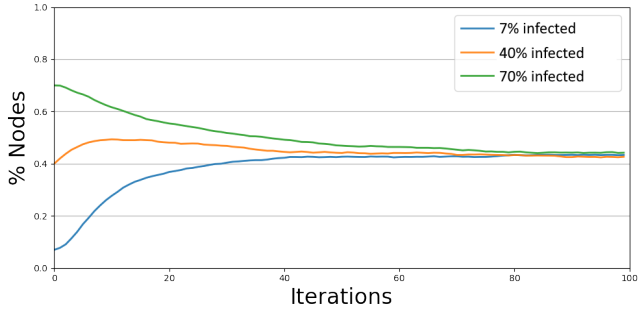

Figure 26: Second Model, Third phase

Also here, no matter what, the infection continue to converge towards a certain equilibrium set around 40%. Similarly, we ran three other batches of tests on networks obtained selecting, respectively, only nodes with weight > 2, weight > 4 and cross-posts >300 [16]. The results weren't significantly different. These results shocked us a lot because theoretically we were expected a pandemic diffusion due to:

$$\lambda = \frac{< k >}{< k^2 >}$$

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?

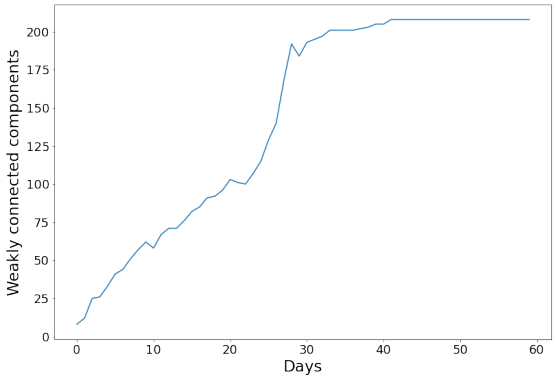

Figure 27: infection throughout the weakly connected components

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 a 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 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.

[16]parameters selected in order to have smaller network

8 DISCUSSION

Our work follows a long process started when we decided to build a network observing the reddit's environment. What we have done was to consider subreddits our nodes and observe their interactions following the scheme of cross-post. We found 24819 subreddits linked by 107094 edges. Then, we characterize our network and confronted it with some theoretical models: Erdos-Reny, Configurational, Watts-Strogatz and Balbarasi-Albert. Confronting all these models, we were able to understand that our network follow a power-low degree distibution and has hubs. Nonetheless, with Leiden method, discover a presence of 45 communities, that have all small dimensions. Moving towards a networks dynamics, we observed the spreading of an epidemics (in our case a post) and predict the links that could be formed within our network. What emerges in spreading was that our network is not easy to infect. That's probably due to the conformation of our network. Link prediction was done following the job of Leiben-Nowel and Keingber[17]. Unfortunately this approach doesn't give to us a particular results, because internet is unpredictable and some posts in the core were to old. Yet, we obtain a better improve in link prediction using some supervised approach and with the K-nearest-neighbour classifier we were able to obtain a prediction of growth around 27%. Of course is still quite small but more significant then the results obtained before. In the last part, following some questions that we have in mind, we observe which subreddits produce more original posts. On the other we observe subreddits with more cross-post. These parts gave us an interesting picture, in fact, it is clear that big subreddits are more likely to produce original content that obtain pretty high up-votes. These content in a second time are posted in small subreddit. Media share content are the original posts more cross-posted and in general subreddits, re-poster or content creators, doesn't have a particular dominant theme. Re-poster subreddits are more likely to share community type posts instead of the content creators that are more likely to create posts bonded with news or media share. The most surprising observation was analyzing the most contents that have a lot of up-votes. What emerges is that having higher up-votes is not enough for a posts to became viral. In fact the posts with more up-votes are cross-posted just in one other subreddit. During the old posts analysis emerges that if a subreddit have a scarce activity, our code has to scroll a lot for reach an "hot" post. At the end, the attention capture by the Game Stop case pushed

[17]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

us to observe how the Game Stop posts were spread around our network because the subreddit considered the starting point is present in our network.Picking all the subreddit that were infected, we compared their infection during the time and the logarithmic price of the GME in the financial market. Surprisingly, the interest on Game Stop case is not correlated with their price, in fact, the majority of subreddit still talk about it also when the price was falling down. Furthermore, the interest practically die when the price return high. We also fitted the data obtained and found a well fit especially during the growth of interest. At the end we ran different kind of SIS simulation to see if we were able to simulate the spreading diffusion of the GME content. For an unexpected reason all the simulation tried, tent to converge into a node infection around 40%. For a deep comprehension we try to observe what could happen to the numbers of weakly connected component once a subreddit is infected. We expected that they remain the same or decrease in time. The contrary happened so new infections do not stem from the infected node neighbors, but pop up randomly, probably caused by the r/all subreddit, the time varying connection and a not complete map of all subreddit present inside reddit. Probably it will emerge something more in further analysis if more and more subreddit will trace and added to our network.

REFERENCES