The Anatomy of a Meme: An Exploration of its Journey on Social Media

Prof. Wenrui Zuo
QMUL

Ashutosh Bhujbal
Student ID - 220166331
QMUL

Jayesh Kumar Soni
Student ID - 220239129
QMUL

Omkar Bare
Student ID - 220459749
QMUL

Shubhankar Shetti
Student ID - 220294807
QMUL

***Abstract*—In general, networks can be used to mimic any kind of spreading process, including the dissemination of ideas, news, and memes, among many other types of spreading processes. For instance, the Susceptible-Infected (SI) model, one of the simplest spreading theories on networks, makes the assumption that each network node might exist in one of two states: susceptible (S) or infected (I). Every node is vulnerable at first, with the exception of one that is in the I state. The likelihood that an infected node will infect its neighbours at each time step is p. This model predicts the long-term evolution, which is the eventual infection of every network node. However, the network's structure has a significant impact on how this is accomplished.**

We may simply expand the SI model in the context of meme spreading by using nodes to represent individuals or groups that "infect" one another by disseminating (sharing) the meme. In this study, we investigate how an imaginary meme spreads over the Reddit hyperlink network, an actual social network. By using simple epidemic spreading models like SI, we seek to examine how the network structure affects the spreading patterns. Do memes swiftly circulate over the entire network? Or perhaps a small community or cluster of nodes has formed as a result of the spreading?

***Index Terms*—Memes, Network Analysis, centrality measures, community**

I. Introduction

To understand how a meme spreads on social media and the factors that contribute to its virality. One should ask the question What is a Meme? Memes are a new way to present one's emotions and are popular among young people. In the current digital era, sharing memes on social media has become commonplace. Memes, which are quickly shared and replicated amusing or satirical images, videos, or texts over social networks, have proliferated into a common form of online communication. It is intriguing from a sociocultural point of view to understand how memes travel through social networks and has real- world implications for business, politics, and public health communication. This study examines how a fictitious meme spreads on the Reddit hyperlink network. A well-known social media site called Reddit is made up of numerous interconnected communities (subreddits) with different levels of user involvement and network structure. We will look into how the network topology affects the spreading patterns of the meme using a straightforward epidemic-spreading model called the Susceptible-Infected (SI) model. We will specifically investigate if the meme quickly spreads over the entire network or if it is mostly focused on particular node communities or small clusters.

II. Related Work

The term "meme" was coined by Richard Dawkins in 1976 to designate "any non-genetic behaviour, going through the population by variation, selection, and retention, competing for the the attention of hosts" [7]. Researchers became aware of a digital version of memes two decades later. In the aftermath of the 9/11 terrorist attacks, sociologists saw a "new genre of cut and paste" jokes that were "parodying, mimicking, and recycling content, embedding it in visual media culture." [8] Multiple people uploaded and modified numerous forms of content, including catchphrases, photos, and video clips, resulting in these new artistic forms of expression. [9] Overall, a meme is a piece of culture, usually with ironic or humorous overtones, that spreads online and develops popularity. [10]

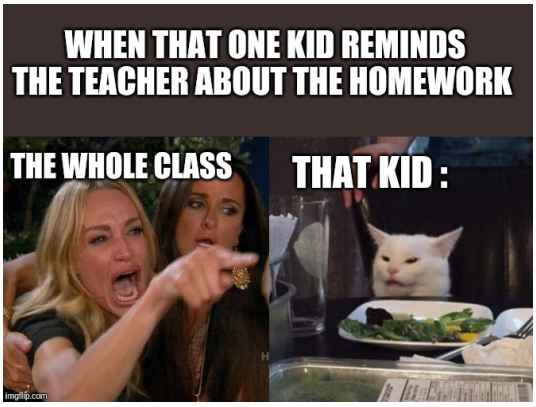

Fig. 1. The viral Woman yelling at a cat meme

Sentiment analysis is a technique used to determine the emotional tone of text data. It has been applied to social media memes analysis to examine users' emotional responses to different types of memes. A study by Chen et al. (2016) [1] analyzed the sentiment of memes shared on Weibo, a Chinese social media platform, found that humorous and emotional memes were more likely to evoke positive emotions in users. Machine learning techniques have been increasingly used to analyze and classify social media memes. A study by Karen Julien et al. (2022) [2] did research how memes can be used as an elicitation tool to help people with interviews. Another

study et al. (2020) [3] used machine learning algorithms to analyze the sentiment (positive, negative, and neutral) of memes, overall emotion (humour, sarcasm, offensive, and motivational) classification of memes, and classifying intensity of meme emotion.

III. Dataset and network presentation

The Reddit Hyperlink Network Dataset is a collection of data from Reddit, a popular social media platform. The dataset contains information about the directed links between different subreddits, which are communities on Reddit. The data was collected over 2.5 years, from January 2014 to April 2017, and is publicly accessible. In total, the dataset contains 35,776 nodes, which represent the subreddits, and 137,821 directed edges, which represent the links between the subreddits. The edges go from the source nodes to the target nodes, indicating the direction of the link. This dataset is useful for analyzing the connections between different subreddits and understanding how information flows through the platform. By studying the Reddit Hyperlink Network Dataset, researchers can gain insights into the structure and dynamics of online communities and how they interact with each other. [4].

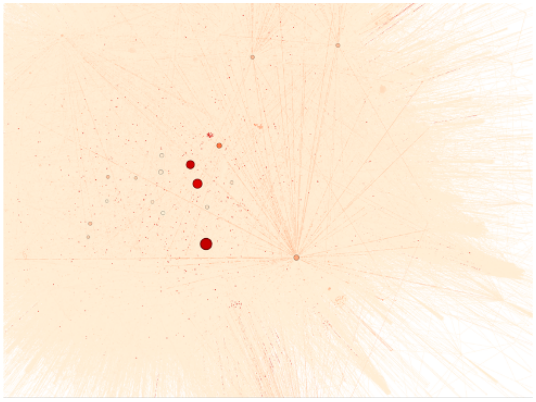

Fig. 2. The graph shows 5-6 nodes with high betweenness centrality.

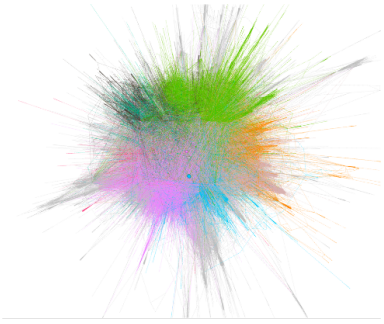

Fig. 3. Statistical Resolutions and Modularity with Partitions

Basic Information About the Dataset –

- Number of Nodes: 35776.
- Number of Edges: 137821
- Average Degree: 3.852.
- Avg. Weighted Degree: 3.852.
- Network Diameter: 13.
- Avg. Path Length: 4.386.
- Avg. Clustering Coefficient: 0.134.

IV. Network analysis methodology

Network analysis is performed on the Reddit Hyperlink (RH) Network The network is extracted from publicly available Reddit data of 2.5 years from Jan 2014 to April 2017." [4].

A. *Task 1*

Under this task, methodologies like data collection and pre-processing were performed which is the first step in the analysis of any model or structure. First, we imported the dataset of edges.csv in the Gephi software tool. The degree to each node with the min and max size was defined to check which node has the highest edges and lowest edges of connections. The force atlas 2 layout was applied to the network, to disperse space around the bigger nodes through the layout method. To prevent the overlapping of nodes and edges in the graph a scaling parameter was set to 50.
A community detection algorithm [5] was used to obtain the communities from the network. Five different "Resolution" values were used to analyse how each value is influencing the resulting community structure (number of communities, the average and standard deviation on the size of communities, among others).

B. *Task 2*

In this project, a SI epidemic spreading model was used to simulate the spreading of a meme in a network consisting of a large volume of nodes and edges. To create this model, the EoN library [6] from Python was utilized, specifically the **fast_SIR()** function. This function allowed for the creation of a SI epidemic model that could be applied to the network to simulate the spread of the meme. Two different centrality measures were considered for this network: degree centrality and closeness centrality. Degree centrality measures the number of edges that a node has, while closeness centrality measures how close a node is to all other nodes in the network. These two measures were chosen because they provide insight into the importance of each node in the network and can help identify key nodes that may be critical for the spread of the meme.

For this task, we have considered the degree centrality and the closeness centrality as the two different centrality measures for this network, which consists of a large volume of nodes and edges.

The tasks were performed 10 times i.e **the number of simulations was 10**. The **transmission rate** was defined to be **0.05**. The **recovery rate** of nodes was set to **0.0** i.e no nodes should be recovered once they are infected.

From the above simulations, a plot of Ti vs Di, where each data point corresponds to a specific node, where Ti is the time steps it took for node i to get infected, and Di is the distance from node i to the initially infected node was generated for each of the highest and lowest centrality measures.

C. *Task 3*

In this task, we analyzed the spreading that occurred within and across the five largest communities that were chosen from Task 1. We used the SI Model to simulate the spread of infection within these communities. For each community, we randomly chose a node to be infected and performed 10 simulations for each community. We plotted the number of infected nodes (Nc) against time for each community and used these plots to analyze the spread of infection within each community. In addition to analyzing the spread of infection within the five largest communities, we also created a random graph using the Python NetworkX Library. This graph had the same number of edges and nodes as the original network. We performed 10 simulations against each of the five largest communities on this random graph as well. We then created plots of Nc against time for each community on the random graph. Overall, this analysis allowed us to gain insights into the spreading that occurred within and across these communities. By comparing the plots for the original network and the random graph using *gnm_random_graph()* we could also analyze the impact of network structure on the spread of infection. This information can be useful for understanding the dynamics of complex networks and designing effective strategies for controlling the spread of infectious diseases or information.

D. *Task 4*

Under this task, we have to remove 5% of the nodes from the original graph. Then, get the 5 largest communities from the graph with 5% fewer nodes. Further, run a SI model on each of those 5 largest communities by infecting a randomly chosen node present in the community. We perform these for 10 simulations on each of the communities. A plot of Nc v/s time is built, where Nc is the number of nodes infected for community c (with c = 1, ...,5). Then the process is repeated by removing 10%, 15%, 20% and 25% of the nodes from the original graph.
Further, the nodes with high eigenvector centrality values are calculated and the same process is performed but instead of removing random nodes from the graph, these nodes are removed.

V. Results and discussion

A. *Task 1*

In order to better understand the structure of the network, a community detection algorithm was utilized. This algorithm allowed us to identify clusters of nodes that are more densely connected to each other than to nodes outside of the cluster. By doing this, we can identify key nodes or clusters within the network and better understand the overall structure and function of the network. To determine the optimal community structure, five different resolution values were tested. Each resolution value influenced the resulting community structure in different ways, impacting the number of communities, the average and standard deviation on the size of communities, and other factors. Ultimately, the partition with a resolution of 1.0 was chosen as it had the highest modularity value of 0.503. Modularity is a measure of the quality of the community structure, and a higher modularity value indicates that the nodes within a community are more densely connected to each other than to nodes outside of the community. In addition to the high modularity value, the partition with a resolution of 1.0 also had a total of 525 communities. This information is useful in understanding the overall structure of the network and identifying key nodes or clusters within the network. By identifying these clusters, we can gain insights into the function of the network and identify areas where improvements or changes can be made to improve network performance. Overall, the use of a community detection algorithm and the analysis of different resolution values allowed for a better understanding of the network and its underlying structure.

Resolution	Modularity	Modularity with resolution	Number of Communities
0.5	0.482	0.225	563
1.0	0.503	0.503	525
1.5	0.459	0.851	753
2.0	0.379	1.225	572
2.5	0.250	1.636	600

We chose the partition with resolution 1.0 as it has better modularity of 0.503, Modularity with a resolution of 0.503 and a number of Communities of 525.

B. *Task 2*

In this analysis, the SI epidemic spreading model was used to simulate the spread of a meme. The researchers used degree and closeness centrality measures, as they were computationally efficient and required less time than other measures. The simulations were performed on the highest and lowest degree and closeness centrality nodes. The node with the highest degree of centrality was identified as 'askreddit'. This node was then selected as the initially infected node, and the SI model simulation was run 10 times. The results of the simulations were plotted on a graph, with Ti on the x-axis and Di on the y- axis. Each data point represented a specific node, with Ti representing the time steps for that node to become infected, and Di representing the distance from the initially infected node. Overall, this study provides insights into the spread of memes and how different nodes in a network may be more susceptible to infection than others. By using the SI model and centrality measures, the researchers were able to efficiently simulate the spread of the meme and identify the node with the highest degree of centrality as the most influential in the spread of the meme. These findings could

have implications for marketing and advertising, as well as understanding how information spreads in online communities.

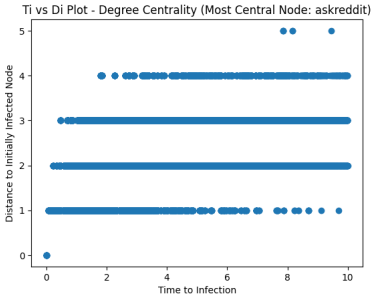

Fig. 4. The graph shows Ti vs Di Plot - Degree Centrality for Most Central Node: *askreddit*

We performed the same analysis by infecting the node with the lowest degree centrality i.e *'ouija irl'*. The following is the Ti vs Di plot for *'ouija irl'*

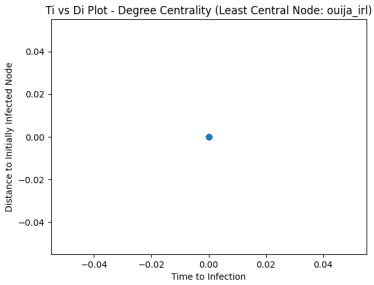

Fig. 5. The graph shows Ti vs Di Plot - Degree Centrality for Least Central Node: *ouija irl*

While for closeness centrality the node with highest closeness centrality was *'askreddit'* while the node with the lowest was *'noshitouija'*.

The Ti vs Di graph are as follows:

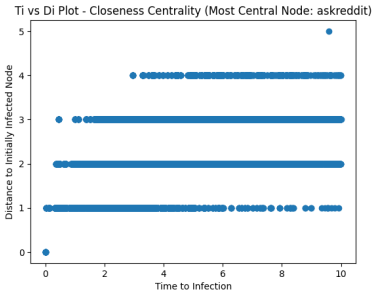

Fig. 6. The graph shows Ti vs Di Plot - Closeness Centrality for Most Central Node: *askreddit*

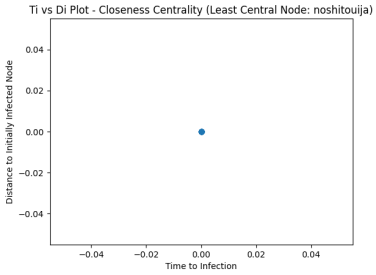

Fig. 7. The graph shows Ti vs Di Plot - Closeness Centrality for Least Central Node: *noshitouija*

C. *Task 3*

In this analysis, the five largest communities in the network were detected using the fast greedy modularity optimization algorithm developed by Clauset, Newman, and Moore [11]. This algorithm is commonly used to identify the communities or clusters in a network based on their connectivity patterns. Once the communities were identified, a SI (susceptible-infected) model was used to analyze the spread of new node infections across the five largest communities.

The SI model is a basic epidemiological model that is commonly used to study the spread of infectious diseases. In this case, the model was applied to the network to simulate the spread of a hypothetical infection. To do this, a node was randomly chosen from each of the five largest communities and the spread of the infection was tracked across the network.

After performing the analysis, the sizes of the five largest communities were determined. Community 1 was found to have 10135 nodes, Community 2 had 8398 nodes, Community 3 had 6135 nodes, Community 4 had 1653 nodes, and Community 5 had 1327 nodes. These results provide valuable insights into the structure and organization of the network.

By identifying the largest communities in the network and analyzing the spread of infections across these communities, we can gain a better understanding of the dynamics of the network. This information can be used to identify key nodes or clusters within the network that may be particularly important for the spread of information or influence. Overall, this analysis provides a valuable tool for studying the structure and function of complex networks and can help us gain insights into a wide range of real-world phenomena. For each of the communities, we have to plot Nc vs Time where Nc is the number of nodes infected for community c with $c = 1,...,5$.

Further, we created a random graph with the same number of nodes and edges as the original network. The same simulations were executed on this graph. The following are the graphs for Community 1. In Fig 8. a random node, (in this case the node name is *proed*) is infected and a SI model is run on the community.

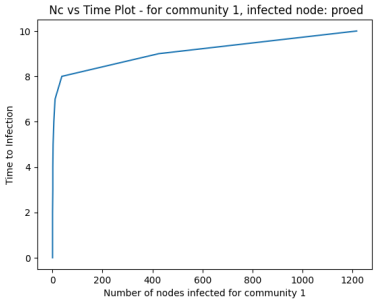

Fig. 8. The graph shows Nc vs Time Plot - For Community 1 Infected Node: *proed*

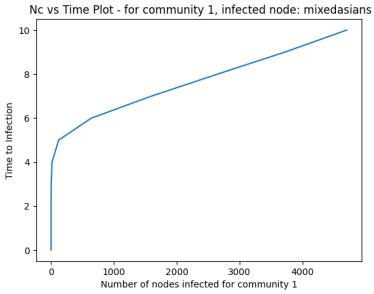

Fig. 9. The graph shows Nc vs Time Plot - For Community 1 Infected Node: *mixedasians*

Comparing the above graphs Fig.8 and Fig 9, In Fig.8, the time taken to infect a node from community 1 took approximately 8 seconds to get from the infected node - proed and within 10 seconds it infected around 1200 more nodes from community 1. In Fig.9, infecting a node from community 1 took approximately 5 seconds to get from the infected node - mixedasians. Within approximately 10 seconds, it infected more than 4000 nodes from community 1. Here we say that the results for Fig. 8 show the infection time taken is more than compared to Fig. 9 where the *gnm_random_graph()* was used to create the random graph and then analyse the difference between the Fig. 8 results.

The closeness of the community can affect the spread of memes on how close a community is to another community. The size of the community can help us analyze the spread of the meme in a community but it is not the only factor which should be considered for analyzing the virality of a meme.

D. Task 4

In this task, we are studying the behaviour of a network after removing a certain percentage of its nodes. Initially, we randomly remove 5% of the nodes from the original network and calculate the five largest communities within it. For each of these communities, we randomly infect nodes and run the SI model to track the number of infected nodes over time. This is done for all five communities, and we plot the results as Nc vs Time, where Nc represents the number of infected nodes in community c. We then repeat the same process by removing 10%, 15%, 20%, and 25% of the nodes from the original network. The aim is to observe the effect of node removal on the spread of infections within the network. By analyzing the results, we can determine how the removal of nodes affects the overall connectivity of the network and its ability to contain the spread of infections. Additionally, we repeat the same process of node removal, but this time, we remove nodes with the highest eigenvector centrality. Eigenvector centrality measures a node's importance within a network based on its connections to other important nodes. By removing nodes with high eigenvector centrality, we can observe how the removal of critical nodes affects the network's connectivity and the spread of infections within it. Overall, the task aims to understand the behaviour of a network after node removal and how it affects the spread of infections. By analyzing the results, we can gain insights into the network's structure and identify critical nodes that play a significant role in its connectivity and overall function. In total, 50 graphs were generated but we have attached only 4 of them for simplicity.

Eigenvector centrality (also called eigen centrality) is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes that themselves have high scores [12].

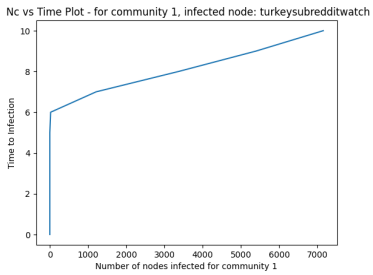

Fig. 10. The graph shows Nc vs Time Plot - For Community 1 Infected Node: *turkeysubredditwatch*

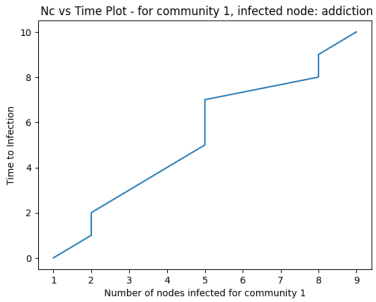

Fig. 11. The graph shows Nc vs Time Plot - without the 10% of nodes with highest eigenvector centrality - For Community 1 Infected Node: *addiction*

The above figures(Fig.10, Fig.11) are a plot of Nc vs time. In Fig.10, A total of 7000 nodes have been infected in a span

of 10 seconds. The first node was infected at 6 seconds. From this, we can observe that most of the nodes got infected in the last 4 seconds. While in Fig.11 a total of just 9 nodes are infected. Although, the first node got infected in 1 second as compared to Fig.10. The difference in the total number of nodes getting affected is due to the Eigenvector centrality measure. Since most of the values with high eigenvector values were removed from the graph for Fig.11. We can clearly observe that it has affected the transmission rate of getting nodes infected in the network.

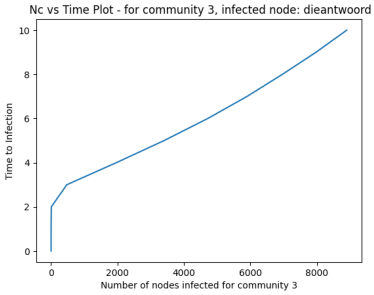

Fig. 12. The graph shows Nc vs Time Plot - For Community 3 Infected Node: *dieantwoord*

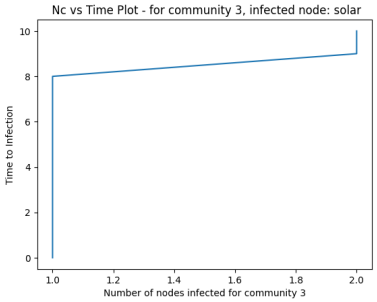

Fig. 13. The graph shows Nc vs Time Plot - without the 25% of nodes with highest eigenvector centrality - For Community 3 Infected Node: *solar*

The above figures(Fig.12, Fig.13) are a plot of Nc vs time. In Fig.12, A total of almost 9000 nodes have been infected in a span of 10 seconds. The first node was infected at 2 seconds. From this, we can observe that most of the nodes got infected in the last 8 seconds. While in Fig.13 a total of just 2 nodes are infected. The first node got infected at the 8^{th} second as compared to Fig.12. The difference in the total number of nodes getting affected is due to the Eigenvector centrality measure. Since most of the values with high eigenvector values were removed from the graph for Fig.13. We can clearly observe that it has affected the transmission rate of getting nodes infected in the network.

Centrality measures come into picture in order to identify the key nodes that are enabling a flow of memes between communities. Few important centrality measures like degree centrality - The degree centrality of a node is simply its degree—the number of edges it has. The higher the degree, the more central the node is. This can be an effective measure, since many nodes with high degrees also have high centrality by other measures. [13]. Another measure is betweeness centrality - The betweenness centrality for each vertex is the number of the shortest paths that pass through the vertex. [14] Closeness centrality - closeness centrality (or closeness) of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph [15].

All the above centrality measures discussed play a significant role in the flow of information through the network.

VI. Conclusions and Perspectives

Our findings provide a overall view of how a meme spreads through a network. It answers questions like How the centrality measures help us understand the network better? how communities interact?. The research has helped to analyze and study how a meme spreads in a network by performing certain analysis like community detection, simulating the spreading of a meme in a network and across communities and identifying key nodes that enable the flow of memes between communities. The tasks involved contributed in comparing the results obtained from different simulations and discussing how different network metrics could be useful in this context. Overall, the research aim was to provide insights into how information spreads in networks and the factors that influence the spread of information.

From a theoretical standpoint, the factors discussed in the paper are valid. However, it is important to consider real-world scenarios where additional factors may come into play, such as the composition of the meme and its sentiment. These factors can significantly impact the spread of the meme as they may influence the level of advocacy and support for the meme. Unfortunately, the current scope of the research does not cover these aspects.

As a future scope, it would be beneficial to investigate how the composition and sentiment of a meme affect its spread. This could involve examining the types of content that tend to be shared and the attitudes that individuals hold towards different types of memes. By exploring these additional factors, we can gain a more comprehensive understanding of the mechanisms behind the spread of memes on social media.

REFERENCES

[1] Chen, Jinyan Becken, Susanne Stantic, Bela. (2018). Sentiment Analytics of Chinese Social Media Posts. 1-7. 10.1145/3227609.3227680.

[2] Julien, Karen. (2022). Using Memes as an Elicitation Tool: The Interview Prompt You Didn't Know You Needed. The Qualitative Report. 10.46743/2160-3715/2022.5640.

[3] [SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!](https://aclanthology.org/2020.semeval-1.99) (Sharma et al., SemEval 2020)

[4] https://snap.stanford.edu/data/soc-RedditHyperlinks.html

[5] Lancichinetti, A., Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Physical Review E, 80(5), 056117.

[6] https://epidemicsonnetworks.readthedocs.io/en/latest/index.html

[7] R. Dawkins et al. The Selfish Gene. Oxford University Press, 1976.

[8] G. Kuipers. Where was king kong when we needed him? The Journal of American Culture, 28(1), 2005.

[9] M. Knobel and C. Lankshear. Online memes, affinities, and cultural production. A New Literacies Sampler, 29, 2007.

[10] P. Davison. The language of internet memes. The Social Media Reader, 2012.

[11] Fast greedy modularity optimization by Clauset, Newman and Moore

[12] https://en.wikipedia.org/wiki/Eigenvector_centrality

[13] Jennifer Golbeck, Chapter 3 - Network Structure and Measures, Editor(s): Jennifer Golbeck, Analyzing the Social Web, Morgan Kaufmann, 2013, Pages 25-44, ISBN 9780124055315. https://doi.org/10.1016/B978-0-12-405531-5.00003-1. (https://www.sciencedirect.com/science/article/pii/B9780124055315000031)

[14] https://en.wikipedia.org/wiki/Betweenness_centrality

[15] https://en.wikipedia.org/wiki/Closeness_centrality