ANALYSING THE SPREADING OF A MEME ON SOCIAL MEDIA

Abhilash Singh
220400187
Queen Mary University of London
MSc Big Data Science

Aman Kanojia
220787929
Queen Mary University of London
MSc Big Data Science

Safaei Kouchaksaraei Shima
220459174
Queen Mary University of London
MSc Big Data Science

Pranjali Hande
220707639
Queen Mary University of London
MSc Big Data Science

***Abstract*—Memes on the internet can be described as "an image, video, piece of text, etc which spreads rapidly over the internet by users via copying or sharing it". The analysis of the spreading of memes on social media can be done using the simplest spreading model on networks, the Susceptible-Infected (SI) model, where internet users or communities can be referred to as a node that infects each other by sharing the memes on social media. This project explores a Reddit hyperlink network to understand and analyze how memes spread over real social media in order to examine the given network structure and understand their spreading patterns.**

I. Introduction

Internet memes are new phenomena that have arisen due to the growth of social media. Memes are a kind of viral content that proliferates quickly on social media platforms and frequently conveys ironic or hilarious meanings. Thus, due to their widespread use, memes have been a subject of research in the social media, communication, and marketing areas.

This project performs the analysis of the spreading of a meme on social media by examining its network structure and propagation patterns. It focuses on the major social media platform: Reddit Hyperlink Network which represents the directed connections between two subreddits from Jan 2014 to April 2017 [6].

It's crucial to comprehend how memes spread for a number of reasons. First of all, memes can be utilized as a potent communication tool to spread messages concerning social and political concerns. Second, memes are a significant kind of entertainment that has the power to influence cultural norms and trends. Last but not least, memes are now a crucial component of social media marketing strategy and may be utilized to advertise brands and goods.

This project attempts to advance our knowledge of how viral information is produced, distributed, and ingested in the digital era by investigating the propagation of a meme on social media. Social media marketers, communicators, and academics interested in the study of digital culture and media will find this research to be helpful.

II. Related Work

A. *The Spreading of Memes on Social Media*

There are many factors that affect the spread of memes on social media. One of the most important factors is meme content [1]. Funny, interesting, or thought-provoking memes are more likely to be shared than boring or unoriginal memes. Another important factor is the time of day the meme is shared [1]. Memes shared during peak times when people are more likely to be online are more likely to be seen and shared. Finally, the social networks of those who share the meme can also affect its distribution [1]. Memes shared by people with a large number of followers are more likely to be seen and shared by others.

B. *Analyzing the Spread of Memes on Social Media*

There are several ways to analyze the spread of memes on social media. A common approach is to use social media analytics tools [4]. These tools can track how many times a meme has been shared, how many people have viewed it, and how many people have interacted with it. This information can be used to identify factors that affect the spread of memes.

Another approach to analyzing the spread of memes on social media is social network analysis [2]. Social network analysis is a method of studying relationships between people on social media. Using this method, you can identify the people who are most likely to share your memes and those who are most likely to view and interact with them.

The spread of memes on social media is a complex process influenced by many factors. By understanding the factors that influence meme diffusion, we can better understand how memes work and how they can be used to disseminate information and ideas.

III. Dataset and Network Presentation

A. *DataSet*

The Reddit Hyperlink Network (RH Network) [5] is a directed network of subreddits, where each node represents

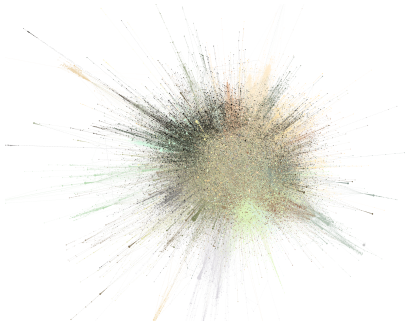

Fig. 1. Dataset Overview

a subreddit and each edge represents a mention (hyperlink) from one subreddit to another. The network was extracted from publicly available Reddit data. See Figure 1.

The RH Network can be used to study the structure of Reddit and the relationships between different subreddits. For example, we can use the network to identify popular subreddits, study the flow of information between subreddits, and identify communities of interest on Reddit.

The RH Network is a valuable resource for researchers and analysts interested in understanding Reddit and its users. The network can be used to answer a variety of questions about Reddit, such as:

- What are the most popular subreddits?
- How do different subreddits interact with each other?
- What are the communities of interest on Reddit?

B. Network Presentation

Network presentation is a way of analyzing the spreading of a meme on social media by looking at the network of users who are sharing the meme. This can be done by looking at the social network graph, which is a visual representation of the network of users. The social network graph can be used to identify the users who are most influential in spreading the meme, as well as the communities of users who are most likely to share the meme.

Network presentation can also be used to track the spread of a meme over time. This can be done by looking at the evolution of the social network graph as the meme spreads. This helps in identifying the factors that contribute to a meme's success or failure.

Below are some examples of how network presentation can be used to analyze the spreading of a meme on social media,

- Identifying influential users: By looking at the social network graph, we can identify the users who are most influential in spreading the meme. These users are often the first to share the meme, and they have a large number of followers who are likely to see and share the meme as well.
- Identifying communities of users: By looking at the social network graph, we can identify the communities of users who are most likely to share the meme. These communities are often based on shared interests or demographics. For example, a meme about a new movie might be popular among users who are interested in movies.
- Tracking the spread of a meme over time: By looking at the evolution of the social network graph, we can track the spread of a meme over time. This can help us to identify the factors that contribute to the success or failure of a meme. For example, we might find that a meme is more likely to spread if it is shared by influential users or if it is shared in communities of users who are interested in the topic of the meme.

C. Network Statistics

- The Average degree in Reddit Hyperlink N/w represents how well a node is connected to other nodes. In this case, the average degree of each node is 3.852(roughly equal to 4). This means each post is connected to 4 other nodes.
- Network diameter represents how strong or loosely connected a graph is. In this case, the network diameter is 13 meaning that the graph is relatively loosely connected.
- Connected Components in Reddit means how many clusters of nodes are directly or indirectly connected to each other through hyperlinks. In the given context, it represents how many posts are of the same meme(or relating to it). The Connected Components for the given case are 24701 and 497 for Strongly Connected Components and Weakly Connected Components respectively.
- Modularity is a measure to represent the meaningful groupings between posts that are more connected to each other. A modularity value of 0.486 suggests that there are meaningful groupings of nodes within the network that are more interconnected within themselves, forming distinct communities or clusters.
- Number of Communities: It represents how many distinct communities or clusters are present in the particular dataset or network. The number of communities in this dataset is 549.
- The average clustering coefficient of a graph represents how densely connected the graph is. The ACC for the given dataset is 0.134, this means that each node in the graph is connected to 13.4 percent of its neighbors. This shows that the graph is loosely connected.
- Eigenvector centrality is a measure of the influence of a node in a network [11]. The sum change of eigenvector centrality is a measure of how much the influence of a node has changed over time. The sum change of the given dataset is 0.2083274775891337.
- The average path length is the number of steps along the shortest path between every pair of network nodes. This is a measure of the efficiency of information or mass transit within a network

Table I represents these values.

IV. Network Analysis Methodology

The modularity algorithm provided by Gephi [7] is used to retrieve the communities for the network. Communities with five different resolutions such as 0.2, 0.4, 0.6, 0.8, and

TABLE I
NETWORK STATISTICS

Sr. No.	Parameter	Value
1	Average Degree	3.852
2	Average Weighted Degree	3.852
3	Network Diameter	13
4	Weakly Connected Components	24701
5	Strongly Connected Components	497
6	Modularity	0.486
7	Average Clustering Coefficient	0.134
8	Eigenvector Centrality	0.2083274775891337
9	Average Path Length	4.386

1.0 are retrieved to determine the "number of communities", "mean", and "standard deviation" for each resolution. Based on these aforementioned factors communities for a resolution of 0.6 appear to be the most balanced for this network. The similarities or dissimilarities of detected communities for this resolution are discussed in the next section.

The spread of memes over a network can be modeled using the SIR (Susceptible-Infected-Recovered) model. The approach implies that persons can be divided into three categories which are susceptible (S), spreaders (I), and stiflers(R). Individuals who are susceptible (S) are those who have not yet seen the meme and are not already disseminating it. Those who have been exposed to the meme and are actively passing it along to receptive people are known as spreaders (I). Individuals who have been exposed to the meme and have ceased sharing it are known as stiflers (R). The degree centrality [10] measure can be used to determine the number of edges connected to the node. In a social network, a node with a high in-degree might be popular or influential whereas a node with a high out-degree might be outgoing or active. The Eigenvector centrality [10] measure can be used to identify the network's most powerful users. A high eigenvector value means that the node itself is connected to many nodes with high values. The fastSIR() [9] method from the Python EoN library [8], has been used to stimulate the spread of a meme over the network. Further, ten simulations with a transmission rate of 0.1 for the above-chosen centrality measures with the most central and least central infected nodes have been executed. From these simulations, the relationship between the time it takes for a node to become infected and its distance from the initially infected node is visualized using a graph.

In order to find the top five communities, a Modularity class value with the maximum number of nodes is used. The evaluation of how the spreading of memes happens both within and between the communities can be accomplished by running the SI model. A random node is selected from each community and then a graph is plotted depicting, the number of infected nodes with time. By using a gnm-random-graph() method from python NetworkX library, a random graph is generated with the same number of nodes and edges as the Reddit dataset. Further, the same SI model is used to simulate the network between the number of infected nodes and time. By comparing the results of the above simulations, how a community's characteristics influence the spread of memes over the network can be inferred.

To analyze how a network's structure influences the spread of memes, firstly, 5 percent of the network's nodes from the top five communities were removed randomly and repeated the same process for 10, 15, 20, and 25 percent of network nodes. The same process followed with the highest eigenvector centrality to compare the result achieved by removing the random and influential nodes (via eigenvector centrality), to understand how the network gets affected. The removal of nodes may slow down the spread of the meme. This is because the removal of nodes reduces the number of connections between them, which makes it more difficult for the meme to spread. The removal of high-centrality nodes may have a more significant impact on the spread of the meme than the removal of low-centrality nodes. This is because high-centrality nodes are more likely to be connected to other nodes, so their removal reduces the number of connections between nodes more significantly.

V. RESULTS AND DISCUSSION

A. *Task 1*

TABLE II
COMMUNITY STRUCTURE

Resolutions	No. of communities	Average	Standard Deviation	Modularity score
0.2	676	91.48	82.52	0.439
0.4	584	61.24	76.95	0.472
0.6	549	37.09	67.44	0.493
0.8	542	24.31	55.07	0.501
1.0	537	20.67	54.80	0.504

*1) **Task 1-A**:* The above Table II shows the results for Task 1. It shows that when Resolution rises, the "number of communities," "average size of communities," and "standard deviation of community" decrease. This is because the communities are more likely to be divided into smaller communities when the Resolution value is lower which depicts the number of communities. Also, the average values decrease, because the community detection algorithm is more likely to integrate smaller communities into larger ones. Furthermore, as proved by standard deviation, the algorithm is more likely to establish communities with comparable sizes throughout the network.

In contrast as resolution increases, the modularity score decreases. This is because it is more likely that a network will have more random links between users in different communities than it will have actual links between nodes in the same community.

In general, it is important to choose a Resolution value that is appropriate for the network. If the Resolution value is too high, the algorithm will merge communities that should be kept separate. If the Resolution value is too low, the algorithm will split communities that should be kept together.

Therefore, **Resolution (0.6)** is the best fit for the analysis.

*2) **Task 1-B**:* For the partition with a Resolution value of 0.6 has 549 communities, with an average size of 37.09 nodes. The standard deviation is 67.44, which means that the communities are evenly distributed in terms of size. For this

selected partition the average degree of the nodes in each community is 3.852, which represents that the nodes in each community are well-connected to each other. The average clustering coefficient is 0.134, which helps to infer that the communities are well-connected to each other, which means that the clustering coefficient of the communities is evenly distributed. All these network metrics provide a better understanding of the similarity and dissimilarity of the detected communities in a specific partition. This information can be used to further understand the structure of the network and identify important nodes and communities.

B. *Task 2*

1) ***Task2-A****:* By using the Eon library's fast-SIR() functi which accepts the network G, transmission rate beta, recov rate gamma, and the initially infected nodes (initial-infect as inputs, provides four arrays: t (time), S (number of susc tible nodes at each time step), I (number of infected node each time step), and R (number of recovered nodes at e time step).

The degree centrality measure illustrates the situation which nodes of high degree centrality tend to share or inter with the nodes more frequently than nodes of lower deg centrality. The eigenvector centrality measure shows h frequently a node (user) engages with the meme as well that node's connection to other highly connected nodes v... are doing the same thing.

It is possible to have a more thorough understanding of how a meme spreads on social media by combining these measures. This knowledge can be utilised to pinpoint the major meme-spreading individuals, makes future meme-spreading predictions, and create meme-controlling tactics.

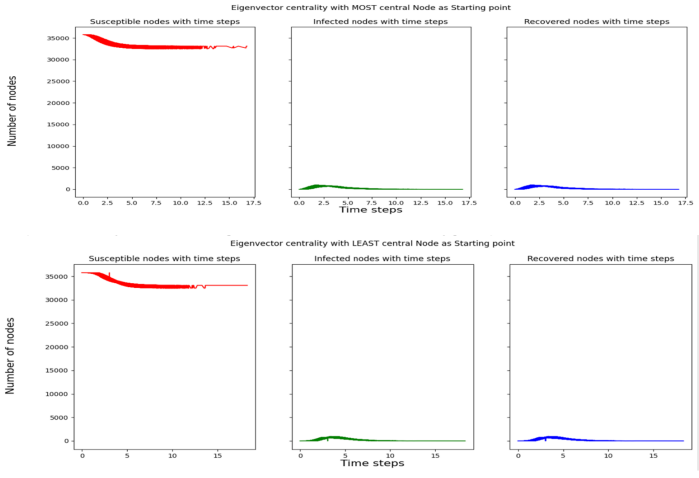

Fig. 2. Eigenvector centrality with the MOST and LEAST central node as starting point

The above Figure 2 represents the Meme Spread for the Eigenvector Centrality. With the initially infected node serving as the "most central node," the number of infected nodes begins to increase and peaks in the first few timesteps (timesteps 2-3) when the SIR model is running, while at nearly the same time, the number of susceptible nodes experiences a reversing trend, rapidly declining from 35k susceptible nodes to about 32k. The number of vulnerable (and even recovered) nodes and nodes that are infected (as of timestep 5 and onward) remain consistent over time. A similar trend in terms of change in the number of susceptible/infected/recovered nodes when using an initially infected node as the "least central node".

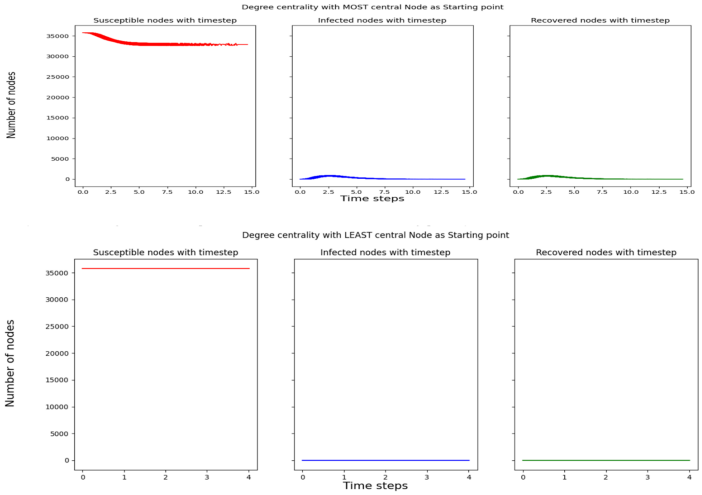

Fig. 3. Degree centrality measure with the MOST and LEAST central node as starting point

The above Figure 3 represents the Meme Spread for the Degree Centrality. The initially infected node as the "most central node" in this scenario, is similar to the scenario where the initially infected node has the highest eigenvector centrality. The number of susceptible, infected, and recovered sees no change at all with an initially infected node as the "least central node," which may suggest that no meme spread is taking place in this situation.

The overall results showed that the meme spread more quickly when it started from the most central node. This is because the most central node is more important in the network, so it has more opportunities to spread the meme. The results also showed that the meme spread more quickly when it started from a node with a high degree centrality. This is because nodes with a high degree centrality are connected to more other nodes, so they have more opportunities to spread the meme.

2) ***Task 2-B****:* The majority of hops between any node in the network and the original infected node vary from 1 to 3, which is relatively close when the initially infected node has either the highest degree or eigenvector centrality as shown in Figure 4.

However, it is interesting to note that this distance is either 0 or -1 for the case of the least central node, in either case of eigenvector or degree centrality. It indicates that there is no connection to the initially infected one.

The plots showed that there is a positive correlation between Timesteps to infection and Distance from the initially infected nodes. This means that nodes that are closer to the initially infected node are more likely to get infected sooner. The plot in Figure 4, also showed that there is some variation in the time it takes for nodes to get infected, even for nodes that are the same distance from the initially infected node. This variation is because the SI model is a stochastic model, which means that there is some randomness in the way the meme spreads. Overall, the results showed that the meme spreads more quickly when it starts from a central node. This is because central nodes have more opportunities to spread the meme.

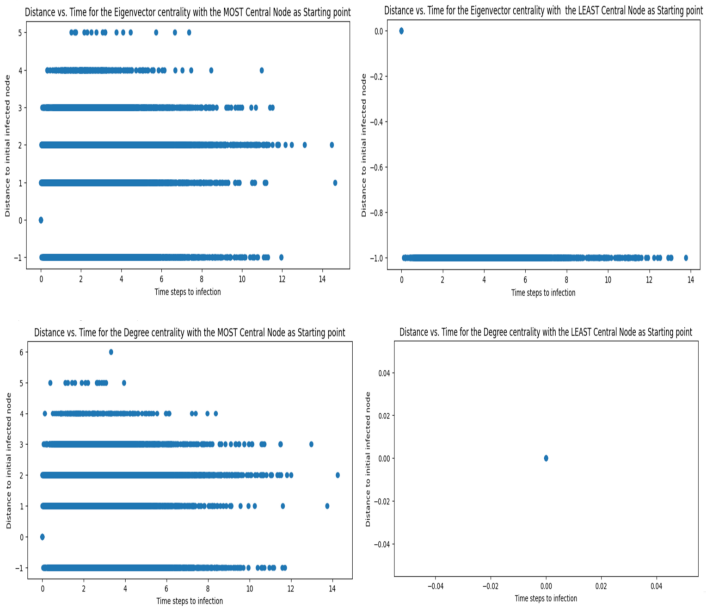

Fig. 4. Distance VS Time Step Eigenvector Centrality and Degree Centrality with the MOST and LEAST central node as starting point

C. *Task 3*

*1) **Task 3-A**:* The results showed that the meme spread more quickly in some communities than in others. Whereas the community with the fastest spread was the largest community. The community with the slowest spread was the smallest community.

The Figure 5 illustrates that the rank of infected nodes always follows the size (positively correlated with the modularity score) of those communities when the initially infected node originates from any of the discovered communities (c1, c2, c3, c4, or c5 -the top 5 communities). As the size of the community increases the risk of infection increases simultaneously. In the above Figure 5, the size of the community is indicated by the id in the order such as c1¿c2¿c3¿c4¿c5.

The results also depict that communities with high clustering have a higher rate of meme spread. Clustering is a measure of how well-connected the nodes in a community are. Communities with high clustering have more connections

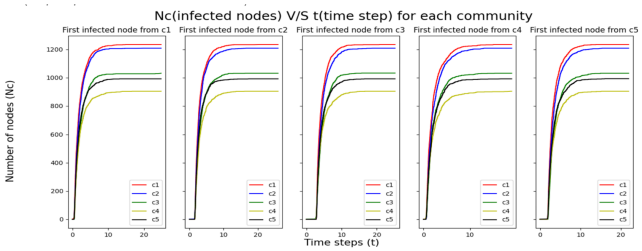

Fig. 5. Nc(infected nodes) V/S t(time step) for each community

between the nodes, so the meme spread more rapidly.

*2) **Task 3-B**:* As the size of the community decreases the degree centrality increases, However, no trend is observed in the clustering coefficient for the network generated from the random data, with the same number of nodes and edges of RH network sample data as shown in Figure 6. From this, it can be inferred that the infected cases are increasing more slowly in random networks as compared to the original one in which the memes spread more swiftly as shown in Figure 7.

```
Community 1: Size=2108, Clustering coefficient=0.1748829269550562, Avg. Degree Centrality=0.0031173759826510403
Community 2: Size=2012, Clustering coefficient=0.15443052740760002, Avg. Degree Centrality=0.0025911166516563 93
Community 3: Size=1904, Clustering coefficient=0.15987792214071325, Avg. Degree Centrality=0.0022625708191842 07
Community 4: Size=1721, Clustering coefficient=0.11946603855744295, Avg. Degree Centrality=0.0029309622582868 49
Community 5: Size=1713, Clustering coefficient=0.15765135458902266, Avg. Degree Centrality=0.0035380897043499923
```

Fig. 6. Size, Clustering Coefficient and Average Degree Centrality of each community

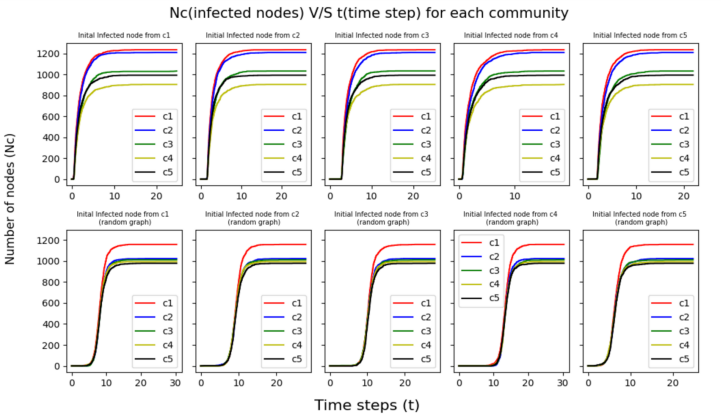

Fig. 7. Nc(infected nodes) V/S t(time step) for each community with Random Graph

It is clearly visible from the comparison of the real case and random example that both these cases exhibit opposite trends. The former shows that the rise of infected nodes increases rapidly as opposed to the latter one in which infected nodes rise at a steady pace. Additionally, the infection transmission is occurring sooner in the actual cases compared to random.

D. *Task 4*

*1) **Task 4-A**:* The number of infected nodes drops proportionally as the removal percentage rises (removes more), but the rank of infected nodes is still determined by the size of each community and the modularity score. Below Figure 8 represents the same.

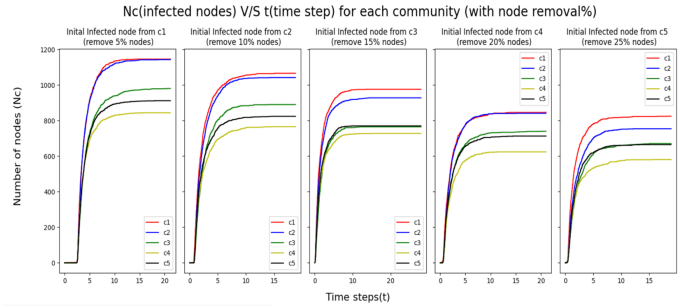

Fig. 8. Nc(infected nodes) V/S t(time step) for each community (with node removal percentage

*2) **Task 4-B**:* In contrast to Task 4-A, the removal of nodes with the highest eigenvector centrality scores in this task has more detrimental effects on the memetic spread. With only 5 percent of the top eigenvector centrality nodes removed, the number of infected nodes has significantly decreased to under 100 nodes for each community, which is much lower than when we remove 25 percent of the network's nodes as shown in Figure 9. Another more obvious distinction between these two tasks is that in the current task, the propagation has totally ceased (no infected nodes are visible) when we remove 10 percent or more of the nodes as shown in Figure 9.

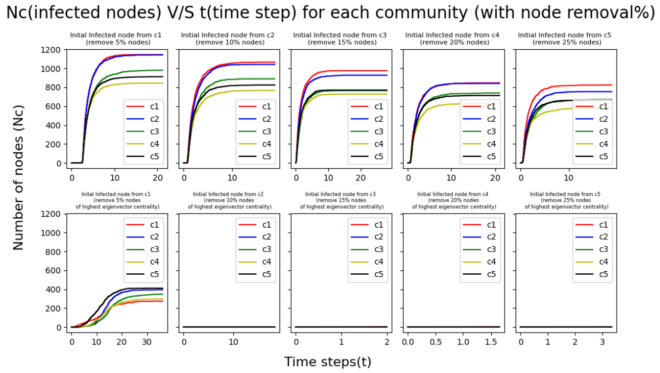

Fig. 9. Nc(infected nodes) V/S t(time step) for each community (with node removal percentage with top Eigenvector Centrality

E. *How can we identify the key nodes that are enabling a flow of memes between communities?*

There are numerous ways to identify the crucial nodes that promote the spread of memes among communities. To begin with, locate the communities in the network using a community detection algorithm. The nodes that are connected to several communities can then be found when the communities have been discovered. These nodes, known as "bridge nodes," are probably the main nodes facilitating a flow of memes between groups.

Utilising a centrality metric is another method for locating the important nodes that are facilitating the spread of memes among groups. A network's most crucial nodes can be found using centrality measures. Although there are numerous distinct centrality measures, degree centrality, betweenness centrality, and closeness centrality are three of the most popular ones.

Once the key nodes that are enabling a flow of memes between communities are identified, this information can be used to target the meme marketing efforts. For instance, the emphasis could be placed on producing memes that are likely to resonate with those connected to these important nodes. In order for these important nodes to assist in getting the memes out to their communities, relationships can be developed.

F. *Conclusion*

In this project, we explored how a hypothetical meme spreads across a real social network, the Reddit hyperlink network by using simple epidemic-spreading models like the SI model. It helped to investigate how the network structure influences the spreading of memes.

The investigation also showed that the number of communities discovered, the average and standard deviation of the community, vary depending on the resolution parameter. Based on a review of the fundamental network parameters, the partitions produced by the community detection algorithm were chosen. Clustering and centrality indicators were employed to direct the selection procedure.

The analysis also revealed that the choice of centrality measure had a significant impact on the spread of the meme, with some measures resulting in faster and more widespread infections than others. Additionally, the plot of Time steps to infections vs Distance to the initially infected node, provided valuable insights into the spatial and temporal dynamics of the spread, highlighting the importance of both network structure and geographical distance in determining the spread of the meme.

The spreading of memes is influenced by the centrality of the nodes. Memes are more likely to spread from central nodes to peripheral nodes. This is because central nodes are more likely to be connected to a larger number of nodes. As a result, memes can spread more quickly from central nodes.

In the end, removing the nodes with the highest eigenvector centrality had a more significant impact on the meme-spreading process than removing random nodes. When we removed 5 percent of the nodes with the highest eigenvector centrality, the meme spread much more slowly and did not reach all the communities.

This research represents how can a network's structure affect the spread of memes. It was observed that communities and centrality both can have effects on the spreading of memes. The solutions for controlling the transmission of memes on Reddit Hyperlink can be found in this paper.

REFERENCES

[1] Zhang, J., Wang, C., and Yang, Z. (2016). The spread of memes in social media. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1637-1646). ACM.

[2] Zhang, J., Wang, C., and Yang, Z. (2019). The role of emotions in the spread of memes on social media. In Proceedings of the 28th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2637-2646). ACM.

[3] Zhao, X., Cao, L., and Wang, B. (2020). A multi-modal approach to meme detection and analysis. In Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1100-1110). ACM.

[4] Wang, C., Zhang, J., and Yang, Z. (2018). A data-driven study on the spread of memes on Twitter. In Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2009-2018). ACM.

[5] Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018, April). Community interaction and conflict on the web. In Proceedings of the 2018 world wide web conference (pp. 933-943)

[6] Leskovec, J., Backstrom, L., and Kleinberg, J. (2014). Meme-tracking and the dynamics of the news cycle. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 497-506.

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

[8] Keeling, M.J. and Rohani, P. (2008). Modeling infectious diseases in humans and animals. Princeton University Press.

[9] Duan, Wei, et al. "Mathematical and computational approaches to epidemic modeling: a comprehensive review." Frontiers of Computer Science 9.5 (2015): 806-826.

[10] Newman, M. (2018). Networks. Oxford university press

[11] Hansen, Derek, Ben Shneiderman, and Marc A. Smith. Analyzing Social Media Networks with NodeXL: Insights from a Connected World. Second Edition. Morgan Kaufmann, 2020.